SIMPLE 3-STEP CENSORED QUANTILE REGRESSION AND EXTRAMARITAL AFFAIRS

Victor Chernozhukov, MIT
Han Hong, Princeton University

Working Paper 01-20
March 2001

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

This paper can be downloaded without charge from the Social Science Research Network Paper Collection at http://papers.ssrn.com/paper.taf?abstract_id=272499
SIMPLE 3-STEP CENSORED QUANTILE REGRESSION AND EXTRAMARITAL AFFAIRS

Victor Chernozhukov, MIT
Han Hong, Princeton University

Working Paper 01-20
March 2001

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

This paper can be downloaded without charge from the Social Science Research Network Paper Collection at http://papers.ssrn.com/paper.taf?abstract_id=272499
SIMPLE 3-STEP CENSORED QUANTILE REGRESSION AND EXTRAMARITAL AFFAIRS

VICTOR CHERNOZHUKOV AND HAN HONG

ABSTRACT. This paper suggests simple 3 and 4-step estimators for censored quantile regression models with an envelope or a separation restriction on the censoring probability. The estimators are theoretically attractive (asymptotically as efficient as the celebrated Powell’s censored least absolute deviation estimator). At the same time, they are conceptually simple and have trivial computational expenses. They are especially useful in samples of small size or models with many regressors, with desirable finite sample properties and small bias. The envelope restriction costs a small reduction of generality relative to the canonical censored regression quantile model, yet its main plausible features remain intact. The estimator can also be used to estimate a large class of traditional models, including normal Amemiya-Tobin model and many accelerated failure and proportional hazard models. The main empirical example involves a very large data-set on extramarital affairs, with high 68% censoring. We estimate 45% – 90% conditional quantiles. Effects of covariates are not representable as location-shifts. Less religious women, with fewer children, and higher status, tend to engage into the matters relatively more than their opposites, especially at the extremes. Marriage longevity effect is positive at moderately high quantiles and negative at high quantiles. Education and marriage happiness effects are negative, especially at the extremes. We also briefly consider the survival quantile regression on the Stanford heart transplant data. We estimate the age and prior surgery effects across survival quantiles.

1. INTRODUCTION

In statistics, biostatistics, and econometrics, there has been a great deal of attention given to censored data. This paper analyzes censored quantile regression models with known censoring points, suggesting a very simple 3-step estimation procedure with congenial features. This simplicity is achieved through the structured envelope and separation restrictions on the censoring probability, yielding an easily implementable and well-behaved technique. These restrictions preserve the plausible semi-parametric, distribution-free and heteroscedastic features of the model. We illustrate the procedure in two examples, an extramarital affairs example and the Stanford heart transplant data with complete censoring times.

The paper is organized as follows. We first evaluate the censored quantile regression model, first proposed by Powell (1986), as a generalization of the Lehman-Doksum p-sample quantile

Key words and phrases. Median Regression, Quantile Regression, Fixed Censoring, Robustness, Accelerated Failure Time Model, Proportional Hazard model, Classification, Discriminant Analysis.

Chernozhukov is at Department of Economics, MIT, vchern@mit.edu; Hong is at Department of Economics, Princeton and UCL, Louvain-la-Neuve, doubleh@princeton.edu.

We thank Takeshi Amemiya, Roger Koenker, and Stephen Portnoy for very constructive, critical comments. We also thank Joshua Angrist, Jerry Hausman, Jinyong Hahn, Sendhil Mullainathan, Whitney Newey for very helpful conversations. Of course, the usual disclaimers apply.
treatment problem for the case of censored data and general treatment. CQR models are semi-parametric, with distribution-free character, and equivariance to monotone transformations. They embody rich information about shifts in location, scale, and other moments induced by covariates, all summarized as the quantile treatment effects. Ensembles of quantile regression curves provide a more complete, often more interesting picture of relationships in the data than conventional mean or median regressions alone. Graphically a spectacular technique, they rouse curiosity, and engage attention. CQR models compare favorably to the well-known Amemiya-Tobin, Cox, Buckley-James, and other approaches, as they permit distribution-free specifications as well as rich forms of heteroscedasticity, including scale and non-scale forms. We briefly discuss available estimators for the cases of fixed censoring. This discussion motivates the k-step estimators on congeniality grounds, including implementation ease, good performance in small samples and models of many continuous or discrete regressors, and cases of very high censoring. These qualities especially pertain the empirical examples, which follow after the discussion of theoretical properties and simulations. The k-step CQR offers a constructive, robust, and well-behaved method to estimate the CQR models as well as the traditional Amemiya-Tobin, accelerated failure, and many Cox models.

2. Censored Regression Quantile Models

2.1. Censored Quantile Regression and Quantile Treatment Effects. Although the need for conditional quantile models had been recognized early in the 19th century, only in the past century did an extensive research of the subject begin. Lehmann (1974) and Doksum (1974) formulated the theory of quantile-quantile plots, posing a p-sample quantile treatment problem and arguing that location-shift models are insufficient to summarize ubiquitous quantile shift effects. Koenker and Bassett (1978) introduced quantile regression (QR) estimators that have evolved into a popular approach to data analysis. Another early work of high merit by Hogg (1975) suggested instrumental variable estimators and gave a first empirical illustration of the conditional quantile model’s breadth. A number of seminal works also lay the foundation of the initial developments, including Amemiya (1981), Chaudhuri (1991), Chaudhuri, Doksum, and Samarov (1997), Powell (1986), Koenker and Portnoy (1987), Portnoy (1991), Jureckova and Prochazka (1994), Newey and Powell (1990), Buchinsky and Hahn (1998) Koenker and Machado (1999), Khan and Powell (2001), as well as many others. QR is central to a substantial number of empirical studies, as Koenker and Hallock (2001) recently reviewed.¹

Our target is the conditional quantile function of the dependent real variable Y given covariates X in \( \mathbb{R}^d \), \( Q_{Y|X} \). \( Q_{Y|X} \) is the inverse of the conditional distribution function \( F_{Y|X} \):

\[
Q_{Y|X}(\tau) = \inf_{v \in \mathbb{R}} \{ v : F_{Y|X}(v) > \tau \};
\]

therefore \( Q_{Y|X} \) is a complete description of the stochastic relation of Y to X.

The linear model of \( Q_{Y|X} \) is of fundamental importance. It is convenient, conceptually appealing and simple, incorporating classical linear model and linear location-scale models

¹Koenker and Gelling (2001)’s introduction to quantile regression is very worthy.
as important special cases,
\[ Q_{Y|X}(\tau) = X'\beta(\tau). \] (1)
We assume that \( X \) includes a constant and note that it may incorporate a wide array of polynomial and alternative transformations of the observable covariates. This model could be specified for a particular quantile, say median, or for an ensemble of quantile curves, providing a more complete, global description of conditional distribution. Both local, or single quantile restriction models and global models each have their own virtues, which we shall stress en-route. In sequel we assume that (1) applies to the quantiles of interest. Linear QR model (1) is a natural generalization of the classical location-shift model:
\[ Y = X'\alpha + u, \] (2)
where \( u \) is independent of \( X \), with distribution function \( F \), and of the linear location-scale shift model
\[ Y = X'\alpha + X'\gamma u, \] (3)
where \( X'\gamma > 0 \). Indeed, in the first case \( \beta(\tau) = \alpha + F^{-1}(u)e_1 \), where \( e_1 = (1,0,...)' \), and in the second \( \beta(\tau) = \alpha + \gamma F^{-1}(u) \). Thus, (2) implies that all the slope coefficients are the same for all \( \tau \), whence (3) implies that \( \beta(\tau) \) are monotone in the quantile index \( \tau \). The general QR model (1), local or global, does not make such restrictions. Thus, QR models ascribe a rich role to covariates \( X \), allowing them to exhort location, scale, kurtosis, and quantile-shift effects. As is the case with conditional mean models, the QR models are distribution free.

Another principal view of QR is as a way to extend the Lehmann-Doksum p-sample quantile treatment problem to the regression setting(Koenker and Gelling (2001)). In the two-sample setting, Lehmann (1974) and Doksum (1974) arrived at the following formulation. Suppose where the value of the untreated is \( y \), the treatment adds amount \( \Delta(y) \). If \( Y \) has the untreated distribution \( F \), then the treated distribution, \( G \), is defined by the random variable \( Y + \Delta(Y) \). Thus, Lehmann defined the treatment effect \( \Delta(y) \) as the horizontal distance between \( G \) and \( F \) in \( (y,p) \) coordinates \( F(y) = G(y + \Delta(y)) \) or
\[ \Delta(y) \equiv G^{-1}(F(y)) - y. \]
Evaluating \( \Delta(y) \) at a quantile, \( y = F^{-1}(\tau) \), one obtains the quantile treatment effect:
\[ \delta(\tau) = G^{-1}(\tau) - F^{-1}(\tau); \]
In fact, \( \delta(\tau) \) is the familiar distance from the 45 degree line in the quantile-quantile plot of \( F^{-1}(\tau) \) vs. \( G^{-1}(\tau) \). For example, the quantile treatment effect could take a simple location-shift form \( \delta(\tau) = \delta_0 \), for all \( \tau \), or a scale-shift form \( \delta_0(\tau) = \delta_0 + \delta_1 F^{-1}(\tau) \). More generally, the treatment can affect such features as skewness, kurtosis, and other moments, all summarized by the quantile shift effect \( \delta(\cdot) \). This 2-sample formulation leads to the linear model (1), noting that \( Q_{Y|D_i}(\tau) = F^{-1}(\tau) + \delta(\tau)D_i \), where \( D_i \) is the treatment indicator.

Thus, firstly, we may view linear QR models \( X'\beta(\tau) \) as a generalization of the \( p \)-sample problem to the case of continuous or polychotomous treatment, represented by covariates \( X_{(j)}, j \geq 2 \), as in dose-response studies. In this case, \( \beta_{(j)}(\tau) \) can be interpreted as a partial derivative with respect to change in the treatment, or, equivalently, a quantile treatment...
effect of changing the treatment \( X_{(i)} = x_0 \) to \( x_0 + 1 \). Secondly, introduction of covariates \( X \) is a pertinent way to control for observed heterogeneity. Regardless of whether covariates \( X \) bear causal or control meanings, we refer to \( \beta(\tau) \) as the quantile treatment or the quantile shift effects.

Equivariance to monotone transformations is a useful property of quantile regression models, as emphasized in Powell (1991), Chaudhuri, Doksum, and Samarov (1997), and Koenker and Gelling (2001) in connection to censoring, survival, and transformation models. We shall state a slightly more general form. For a given measurable transformation \( T_z(Y) \) of variable \( Y \) and other variables \( Z \), it is obvious that

\[
Q_{T_z(Y) \mid X, Z}(\tau) = T_z(Q_{Y \mid X, Z}(\tau)), \quad \text{since}
\]

\[
P[Y \leq Q_{Y \mid X, Z}(\tau) \mid X, Z] = P[T_z(Y) \leq T_z(Q_{Y \mid X, Z}(\tau)) \mid X, Z].
\]

For example, if we estimate a linear model \( X'\beta(\tau) \) for the logarithm of survival time \( T \), \( Y = \log(T) \), as in accelerated failure time models, then \( Q_T(\tau \mid X) \) equals \( \exp(X'\beta(\tau)) \). This property helps interpret and communicate data-analytic findings; and it is not shared by the conditional mean models. Koenker and Gelling (2001) contains a lucid illustration in the case of quantile regression survival analysis.

Transformation equivariance naturally leads to models of censored data. Assume that the latent variable \( Y_i^* \) is left-censored\(^2\) by the observable, possibly random, censoring points \( C_i \), and we collect

\[
Y_i = Y_i^* \vee C_i, \quad X_i, \quad \delta_i = 1(Y_i = C_i).
\]

Assume \( Y_i^* \) is conditionally independent of the censoring point \( C_i \), that is, for all \( y \in \mathbb{R} \):

\[
P(Y^* < y \mid X_i, C_i) = P(Y^* < y \mid X_i), \quad \text{so that}
\]

\[
Q_{Y^* \mid X, C_i} = X'\beta(\tau).
\]

The conditional independence assumption is more realistic than the frequently made assumption of independence between \((Y_i, X_i)\) and the censoring variables \( C_i \). (In the Stanford example, for example, there is a significant correlation between the "acc" and "surgery" variables with the censoring times). The assumption, that censoring points are known for all \( i \), is realistic in many (but clearly not all) situations. For example, in the analysis of post-transplant survival times in the Stanford data-set, we can compute all censoring points because we know the transplant and the last follow up dates for each \( i \). In other situations, it is possible to impute them (see Powell (1986) for discussion).\(^3\) In the extramarital affairs example, the censoring point is, naturally, zero, or "fixed", for all observations. This type of censoring is very common in social, psychometric, technometric, and econometric studies. Conditioning on \( C_i \), assumption (6) and transformation equivariance yield the following censored QR model:

\[
Q_{Y \mid X, C_i}(\tau) = X'\beta(\tau) \vee C_i.
\]

\(^2\)right-censoring is handled by reversing the sign.

\(^3\)Here we do not consider the cases where the censoring points are unobserved. There is a growing literature on median and M-estimation under random censoring – see e.g. Yang (1997) and Zhou (1992).
Early formulations of (7) go back to Amemiya (1973) and Tobin (1958) who considered Gaussian errors and fixed censoring point at 0:

$$Y_i = X'\beta + u, \text{ if } X'\beta + u > 0 \text{ and } 0, \text{ if not } ; \quad u \sim N(0, \sigma); \quad (8)$$

This model states that $Q_{Y|X}(1/2) = X'\beta \lor 0$ and assumes a parametric form of distribution of error $Y_i - X'\beta \lor 0$.

Historically, Tobin (1958) first formulated (8) and Amemiya (1973) provided and justified consistent estimators. Analyzing expenditure on durable goods, Tobin accounted for the fact that expenditure was nonnegative and frequently assumed value 0. In many examples, specifically Tobin’s example, such censoring formulation is recurrently criticized, since the negative values are sometimes meaningless or given imaginary interpretations. We stress that although $x'\beta$ may be hard to interpret as a mean of a ”latent variable,” the conditional median model $Q_{Y|X}(1/2) = X'\beta \lor 0$ for the observed expenditure is an excellent one. Indeed, if conditional on $X$ probability of ”censoring” is greater than 1/2, then conditional median of expenditure is zero. Otherwise, the conditional median is (reasonably approximated as) $X'\beta$.

In his path-breaking work, Powell (1984) was first to stress median and quantile regression in the distribution-free, fixed censoring world. For this reason, we refer to this model as to the Powell CQR model, and to its normal form as to the Amemiya-Tobin model.

Having model (7) in mind, we can speak of the quantile treatment/shift effects. $\beta(\tau)$ now summarizes the effect of a treatment on both the censored and uncensored latent variable. When the latent variable has real meaning, such as the survival time, the interpretation is the same as discussed earlier, except that censoring may prevent identification of the treatment effects for all quantiles of interest. When the latent variable is imaginary, as in the expenditure or extramarital examples, it is important to indicate the relevance of $\beta(\tau)$.

Define the censored quantile treatment/ shift effect as a partial derivative

$$\delta_j(\tau, x, c) \overset{\text{def}}{=} \frac{V_{Y|x+e_j, c}(\tau) - V_{Y|x, c}(\tau)}{V}$$

$$= \frac{((x + e_j v)'\beta(\tau) \lor c - x'\beta(\tau) \lor c)}{V}$$

$$\equiv 1(x'\beta(\tau) \geq c)\beta(\tau)$$

$\delta_j(\tau, x, c)$ characterizes the effect of changes in components of $x$ keeping everything else fixed ($e_j$ is a zero vector with 1 in the $j$-th position). $\beta(\tau)$ in turn characterizes the treatment effect $\delta(\tau, x, c)$.

As the class of CQR models inherits the qualities of the uncensored QR models, such as distribution-free character, talent of heteroskedasticity, they compare favorably to the normal Amemiya-Tobin models, distribution-free homoskedastic accelerated failure models, and Cox proportional hazard models. For example, an accelerated failure time models are of the form

$$\log(T) = \alpha + X'_{-1}\theta + u,$$

where $u$ has an unknown distribution $F$, $X_{-1}$ represents covariates without the intercept. This is a location-shift model with quantile treatment effect given by $\theta$. Many useful proportional hazard Cox models are special cases of (9), for example, one with Weibull baseline hazard model. More generally, Cox models can be written as $\ln \Lambda_0(T) = \alpha + X'_{-1}\theta + \sigma u$, for unknown
integrated baseline hazard function $\Lambda_0$ and a Gumbel variate $u$. Vector $\theta$ summarizes the relevant quantile treatment effects, so that the Cox model is a location-shift one up to a transformation. From a constructive point of view, one or the other model may be preferred depending on the context. However, when heteroscedasticity and distribution free spirit are desired, the CQR approach is valuable.

2.2. Motivation for k-step Estimators. Sample regression quantiles may be defined in several ways. The most widely used method, due to Koenker and Bassett (1978), is ingeniously simple. Suppose we have $n$ observations $\{Y_i, X_i\}$. In the no-covariates case, the sample $\tau$-th quantile $\hat{\beta}(\tau)$ is generated by solving the problem (Ferguson (1967)):

$$\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} \rho_{\tau}(Y_i - \beta),$$

where $\rho_{\tau}(x) \equiv (\tau - 1(x \leq 0))x$. Koenker and Bassett (1978) extended this concept to the regression setting as follows:

$$\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} \rho_{\tau}(Y_i - X'_i\beta).$$ (10)

They also and Powell (1984) and Portnoy (1991) developed the asymptotics. The asymptotic distribution parallels that of ordinary sample quantiles. Also, as Koenker and Bassett (1978) elaborated, the sample regression quantiles inherit good robustness and equivariance features of the ordinary sample quantiles. Robustness is a strong motivation for quantile regression even in canonical, seemingly normal, homoskedastic cases, as heavy-tailed admixture or outliers are known spoilers of many classical procedures.

For brevity assume for now $C_i = 0, \forall i$. In the censored model (7), replacement of the linear form with the semi-linear one ($X'_i\beta \lor 0$),

$$\min_{\beta} \sum_{i=1}^{n} \rho_{\tau}(Y_i - X'_i\beta \lor 0),$$

leads to the celebrated Powell (1984), Powell (1986) estimator. Powell (1984) established the asymptotic normality of $\hat{\beta}_p(\tau)$ and developed an inference theory, paving the way and setting forth a standard for future work.

Despite the intuitive appeal, this method has not become popular in empirical research due to its well known computational difficulty.\footnote{We know of about three or four applications of censored median regression in econometrics. In contrast, Amemiya-Tobin or Cox approaches found hundreds of applications, if not thousands.} See, for instance, Fitzenberger (1997b), Fitzenberger (1997a) for remarkable research as well as extensive simulations. Buchinsky (1994) and Fitzenberger (1997a) designed ingenious computational algorithms, which Fitzenberger (1997a) recommends for low degrees of censoring, while admitting that "all practical algorithms perform quite poorly when a lot of censoring is present".\footnote{For example, see Fitzenberger (1997a), p.15. In the case of 50% censoring, one regressor, and small sample $n = 100$, in several important designs (e.g. A, B), the frequency of computing the Powell estimator ranged from 5% to 37% for various algorithms. For some of the other designs results were better – convergence
substantiated by a very extensive Monte-Carlo experiment with tens of different practical designs. All experiments involved only one regressor (!). Increase in dimension leads to further complications. In many empirical applications, the censoring is quite heavy and dimensionality is also high. For example, in the affairs example of section (4), degree of censoring is 68%; in the heart dataset of section (3.6.2) – 37%. The number of regressors in these two datasets are 9 and 3, respectively. This serves as one strong but not the only motivation for the matter herein. Arguably, an important goal of statistic theory is the design of both theoretically elegant and also implementable, practically attractive estimators. It is that requirement that makes the problem at hand particularly challenging.

In part motivated by such limitations, recent remarkable work by Buchinsky and Hahn (1998) and Khan and Powell (2001) suggested a number of alternative estimators. Buchinsky and Hahn (1998) proposed to first estimate the propensity score \( h(X_i) = P(\delta_i = 1|X_i) \) by a nonparametric kernel regression, then select a set where \( \{i : h(X_i) > 1 - \tau\} \) of the whole sample, i.e. those observations \( i \) where the conditional quantile line is above zero, \( X'_i \beta(\tau) > 0 \), and then use a transformed QR on the selected sample. Khan and Powell (2001) also proposed to use any of the three methods to perform the first stage selection: (i) maximum score estimators of the regression quantile, (ii) nonparametric kernel propensity score estimator \( \hat{h}(X_i) \), (iii) nonparametric locally linear conditional quantile estimator of Chaudhuri (1991) (denoted \( \hat{q}(X_i) \)). In the second step, they obtain the estimator by running a weighted QR:

\[
\min_\beta \sum_{i=1}^n \Lambda_i \rho_{\tau}(Y_i - X'_i \beta) \quad \text{where e.g.} \quad \Lambda_i = \Lambda(\hat{h}(X_i) - (1 - \tau) - c) \quad \text{or} \quad \Lambda(\hat{q}(X_i) - c), \quad \text{etc.}
\]

The two-stage estimators are somewhat less efficient than the Powell estimator due to smoothing and trimming. Ideologically, these estimators share the ideas behind the construction of the Powell (1986) estimator (especially the estimating equation version of it), except that Powell imposed simultaneity to obtain his single step estimator.

The suggested first stages are extremely attractive, but are only practical in low dimensions, and have slow convergence rates. Local kernel smoothers apply to (sufficiently) continuous variables only, whereas a lot of applications, including ours, have many (sufficiently) discrete covariates. This is very confounding. Of course, an asymptotic theory suggests that the \( \sqrt{n} \)-estimates could be obtained by averaging within the cells. In the affairs example, there will be on average roughly \( 6800/(8^3) \approx .2 \) observations per cell; in the heart example – \( 69/(15 \times 2^2) \approx 1 \). Thus such asymptotics is void here. The computational burden [especially when the bandwidths are carefully chosen] is very substantial in high dimensions, large datasets for all of the first stage estimates. As a result, we simply cannot use any of the available estimators in our examples and many other real-life applications due to 1) heavy censoring, 2) high dimensionality, 3) very small or large sample, and 3) polychotomous nature of numerous regressors. As our interest concerns many quantiles; these settings are even more confounding. From a constructive angle, however, we strongly stress that the mentioned estimators can be potentially fruitful in a lot of conceivable cases.

In an unpublished report that precedes this one, we suggested a number of flexible parametric techniques that exploit the global approximation and classification ideas, that lead

---

frequencies ranged from 30% to 70%. These results were obtained for the case of one regressor. In case of many regressors and larger \( n \), the results can be expected to be worse.
to consistent and asymptotically normal estimators that are as efficient as the benchmark Powell estimator. The present paper extracts and further develops what we think is an essential, most implementable and usable part of our previous research. The current approach is based on the structured modeling restrictions that we put on the censoring probability. These restrictions cause a reduction in generality, but only a small one, since they allow to incorporate, for example, Amemiya-Tobin, many Cox models, and accelerated failure time models as very important special cases while at the same time allowing for heteroscedasticity and distribution-free character. As a result, an easily computable (comparable to linear least squares), well-behaved, robust estimator is available. Due to the good robustness properties, it offers not only an efficient, practical way to estimate the general Powell CQR models, but also a good way to estimate very important traditional models. Because it is easily computable, its finite sample properties can be studied in fine details.

3. Simple 3-step and k-step CQR Estimators

3.1. The procedure: This section describes the steps of the estimator and briefly sketches the basic ideas behind them.

3.1.1. The Steps.

Step 1. Estimate a parametric classification (probability) model:

\[ \delta_i = p(\hat{X}_i'\gamma) + \epsilon_i, \]

where \( \delta_i \) is the indicator of not-censoring. \( \hat{X}_i \) indicates desired transforms of \( X_i \) and \( C_i \). Next, select the sample \( J_0 = \{ i : p(\hat{X}_i'\gamma) > 1 - \tau + c \} \), where \( c \) is strictly between 0 and \( \tau \) and not too small (practical choice is discussed in the appendix).

Step 2. Obtain the initial (inefficient) estimator \( \hat{\beta}_0(\tau) \) by running the standard QR:

\[ \min_{\beta} \sum_{i \in J_0} p_\tau (Y_i - X_i'\beta). \]  \hspace{1cm} (11)

Next select \( J_1 = \{ i : X_i'\hat{\beta}_0(\tau) > C_i + \delta_n \} \). \( \delta_n \) is a small positive number going to zero slowly as \( n \to \infty \) (The formal condition on \( \{\delta_n\} \) is that \( \sqrt{n} \times \delta_n \to \infty \) and \( \delta_n \searrow 0 \)). This step therefore consistently selects the largest subset of \( i \) such that \( X_i'\hat{\beta}(\tau) > C_i \), building up the efficiency of the next step.

Step 3. Run QR (11) with \( J_1 \) in place of \( J_0 \). Denote this 3-step estimator \( \hat{\beta}_1(\tau) \).

Step 4. (Optional) Further repeat step 3 once or finite times, using sample \( J_{I} = \{ i : X_i'\hat{\beta}_{I-1}(\tau) > C_i + \delta_n \} \) in place of \( J_0. \) \([I = 2, 3, \ldots]\).

In step 4 each repetition involves selecting \( J_{I} = \{ i : X_i'\hat{\beta}_{I-1}(\tau) > C_i + \delta_n \} \), and then obtaining \( \hat{\beta}_I(\tau) \) from (11) using the sample \( J_{I} \). Denote the k-step estimators as \( \hat{\beta}_I(\tau) \).

\(^6\)Made for brevity of proofs, the restriction of single index can be relaxed by allowing the binary model to have plausible stochastic complexity as measured by the uniform covering or bracketing entropy. The class of the single index is particularly suited here from a practical point of view.
3.1.2. What is in the steps? Some details are as follows. In the step 1 we may use, for example, logit, probit, extreme value, linear (polynomial) or any other model that fits the data \{\delta_i, X_i\} well. \(\hat{X}_i\) denotes any desired transform of \(X_i\). For example, \(\hat{X}_i\) may consist of \(X_i, C_i\) and its squares (Remark 4 allows for power series and regression spline approximations). In general, this gives an inconsistent estimator of the propensity score

\[ h(X_i) \equiv P(\delta_i = 1|X_i, C_i), \]

but the inconsistency is not important as long as the misspecification is not too severe.

Indeed, the goal of the step 1 is to select some and not necessarily the largest subset of observations where \(h(X_i) > 1 - \tau\), i.e. where the quantile line \(X'_i\beta(\tau)\) is above zero, so as to obtain an initial consistent but inefficient estimator \(\hat{\beta}_0(\tau)\). This task can be carried out if, for example, \(p(\hat{X}_i'\gamma_0) - c\) is a lower bound on \(h(X_i)\):

\[ a.s. \quad p(\hat{X}_i'\gamma_0) - c < h(X_i), \tag{12} \]

\((\gamma_0 \equiv \text{plim} \hat{\gamma})\) and it is nontrivial, meaning that the selected set \(J_0\) is sufficiently rich – matrix \(EX_iX'_i\{i \in J_0\}\) is asymptotically invertible. Greater \(c\) and better model \(p(\cdot)\) simplify the selection task. This is a weak requirement since \(p(\cdot)\) is not required to be a distribution function. Thus the envelope restriction is intuitive and not very restrictive, but it may be replaced by an even weaker condition – the separating hyperplane\(^7\) restriction in Theorem 1 (cf. Figure 1 for motivation). See the subsections below for formal details and further discussion of this assumption. Appendix B contains details for practical implementation.

The above construction assumes that the estimator \(\hat{\gamma}\) is reasonable and converges to a value \(\gamma_0\) that minimizes a sensible distance between \(h(X_i)\) and the model \(p(\hat{X}_i'\gamma)\). For example, \(\hat{\gamma}\) may be defined by minimizing \(\sum_{i=1}^{n}[\delta_i - p(\hat{X}_i'\gamma)]^2\), in which case under standard conditions \(\hat{\gamma}_0 \to \gamma_0\) that solves \(\min_{\gamma} E[h(X_i) - p(\hat{X}_i'\gamma)]^2\). Alternatively, quasi-ML methods can be used and will be equivalent to weighted least-squares. In our empirical section we employed a polynomial logistic model and estimated it using conditional MLE.

Another attractive choice is the Fisher-Rao discriminant analysis. The discriminant prospective is justifiable as follows (treat \(C_i\) as part of \(X\) or set \(C_i = 0\) for brevity). If \(X_i\{|\delta_i = 1\}\) has density or p.m.f. \(g_1(x)\) and \(X_i\{|\delta_i = 0\} - g_0(x)\), then by the Bayes’s rule:

\[ P(\delta_i = 1|X_i = x) = \frac{q_1 g_1(x)}{q_1 g_1(x) + q_0 g_2(x)}, \]

where \(q_1 = P(\delta_i = 1) = 1 - q_0\). Approximating \(g_1(x)\) and \(g_2(x)\) by normality with different means and variances, leads to the classical logistic linear-quadratic discriminant analysis, LQDA (see e.g. Amemiya (1985), p. 282). Other forms of \(g_1\) and \(g_2\) could be chosen in view of their concrete problems, but even normality assumption has been known to produce good results (cf. Efron (1975), Press and Wilson (1978), Amemiya and Powell (1983)). Using both simulated and real-life examples, the studies found a surprisingly good performance of the LDA classifier even when the regressors were binary! Naturally, when the normal approximation becomes realistic, discriminant analysis does much better than conditional approaches (because it approximates the unconditional MLE). LQDA was among the top 3 classifiers for

\(^7\)In the terminology of modern classification analysis.
11 out of 22 datasets in the impressive Statlog project (see Michie and Taylor (ed) (1994)) and out-competed many sophisticated classifiers. The Statlog project tested 23 algorithms on 22 large-scale, commercially important problems (Michie and Taylor (ed) (1994)). For excellent further refinements and generalizations of the discriminant approach see the literature cited in the next subsection.

Formally we shall not confine ourselves to a particular estimator, instead an assumption such as (12) or its generalization will be required to hold.

Under conditions to be stated, \( \hat{\beta}_1 (\tau) \) is asymptotically normal with variance equal to that of the Powell estimator. Thus, starting with a "good" subset of observations, only two recomputations of QR suffice to obtain the Powell-efficient estimator. Estimators \( \hat{\beta}_1 (\tau) \) are also asymptotically normal with variance equal to that of the Powell estimator. What is the asymptotic rationale for considering the fourth step? For \( I \geq 2 \) estimator \( \hat{\beta}_I (\tau) \) has the efficiency structure of the Powell's one step estimator in the sense that the selector in step 3 has \( \sqrt{n} \) rate of convergence and Powell's variance, which is more efficient than the selector \( \hat{\beta}_0 (\tau) \) used in computing \( \hat{\beta}_1 (\tau) \). This efficiency structure is a unique property of the Powell single-step estimator and our 4-or-more step estimators.

The QR iterations on step 4 are somewhat analogous to those in the pioneering and remarkable ILPA algorithm that Buchinsky (1994) designed for the Powell problem. However, going beyond the third or fourth step is not desirable on statistical grounds (not to mention computational reasons), based on our Monte-Carlo experience. Our result show that given the first classification step only two recomputations of quantile regression lead to an efficient estimator (relative to the Powell estimator). (Regarding computational aspects, the faster interior point algorithms, Koenker and Portnoy (1987), may be preferred to linear programming.)

In summary, the estimation procedure has two very distinct features: a very simple, parametric classification first step, and the additional 3rd and the optional further step. The estimators are as efficient as the Powell estimator.

3.2. The Model Beneath: Normative and Agnostic Prospectives. The canonical CQR model in (7), together with the "nontrivial envelope" restriction (5), can be thought of as a model. This additional assumption is a critical ingredient to yield the simplicity. How restrictive is it?

Set \( C_i = 0 \) to simplify notation. A popular normal model assumes that conditional on \( X_i; Y_i \) is conditionally homoscedastic normal. Then propensity score \( h(x) = \phi(x'\alpha) \), for the normal c.d.f. \( \phi \). A significantly more general CQR model can be immediately obtained by simply assuming that \( \phi(x'\gamma_0) - c \) is a nontrivial envelope of an unknown propensity score \( h(x) \), where e.g. \( x = (x, x^2, ...) \). Such assumption imposes neither normality nor conditional homoscedasticity nor a location-scale sub-model. Similarly, if the benchmark model is, say, the Weibull proportional hazard model, as in the section about the heart example, then

---

8see Buchinsky (1994) or Fitzenberger (1997b) details. The basic idea is to start at a value \( \beta(\tau) \), say 0, and then proceed with iterative linear programming computations until convergence is reached. The convergence to Powell's estimator is not guaranteed and can be quite rare; see Fitzenberger (1997b) and earlier discussions. The convergence to a local optimum does not lead to a consistent estimator—cf. Powell (1984), Powell (1986).
Figure 1. How it works. The solid line depicts the conditional quantile function and the propensity score. The propensity score equals to $1 - \tau$ for the value of $X$ s.t. conditional quantile line $X'\beta(\tau)$ equals the censoring point $C_i = 0$. It is above $1 - \tau$ for $X$ s.t. $X\beta(\tau) > 0$, and below - if $X\beta(\tau) < 0$. Once a sample is s.t. $X_i\beta(\tau) > 0$, the conditional quantile function of the uncensored model can be estimated by the usual linear QR. Initial step involves fitting an "envelope" of the propensity score, and selecting all $i$, s.t. $X_i \in [S1, S1']$. Note that the "envelope" is not ideal, but this is irrelevant, since it acts as a good separating hyperplane selecting a sub-set of $i$ s.t. $X_i'\beta(\tau) > 0$. Next Step fits the QR line, which is used to select all $i : X_i \in [S2, S2']$. The Third Step uses the selected sample, which asymptotically gets close to the ideal: $i : X_i \in [0, S2']$. Note that $h(X)$ can cross the $1 - \tau$ line only once at $X'\beta(\tau) = 0$. In dimensions greater than 1, $X \in \mathbb{R}^d$, the crossing points are defined by the hyperplane $X'\beta(\tau) = 0$, which have a zero span in $\mathbb{R}^d$. Thus, in $\mathbb{R}^d$ the crossing points form a singularity as well.
h(x) = g(x'\alpha) for the Gumbel c.d.f. \( g \). A much more flexible CQR model is obtained by assuming that \( g(x'\gamma_0) - c \) is a nontrivial envelope of the propensity score \( h(x) \).

More generally, Theorem 1 replaces the intuitive envelope restriction by a much weaker separation restriction, which requires that once \( p(x'\gamma_0) \) is above a threshold \( c \), \( h(x) > 1 - \tau \). This assumption allows the envelope to be an incorrect lower bound of the propensity score, but only requires that it does a good job at selecting a correct subset of observations. Figure 1 illustrates the situation. For well behaved models, the further away from zero the conditional quantile function, the further away from \( 1 - \tau \) is the propensity score function, the easier it is to carry out the classification. Classification/discrimination problems of this kind can be subjected to the standard or more modern classification analysis (e.g. Breiman, Friedman, Olshen, and Stone (1984), LeBlanc and Tibshirani (1996), Ripley (1996), Vapnik (2000)). Therefore, in principle, many elaborate, structured strategies for the first classification step are available. We confine our discussion to the simplest, most familiar methods.

The justification outlined above is based on a classical, asymptotic point of view. An alternative is to view such a procedure as a shrinkage method where the first step trades bias of approximating a propensity score envelope for smaller variance. We also argued above that due to the special nature of the problem the bias itself can be quite small. Thus, the shrinkage aspect may be of primary importance in moderate-sized samples. This is confirmed by the computational experiments discussed next.

More generally, from a real-life (agnostic) perspective, either the linear quantile or envelope models are approximations. There is no reason to believe one is more correct than the other, once both are reasonable and flexible models. Therefore, we hope this paper offers an organized way of thinking about and building such models.

In summary, it should be said the model studied here is, of course, somewhat more restrictive than the canonical Powell CQR model. Yet despite being more restrictive, the emphasized model leaves the general, plausible features of the CQR intact. This model is also congenial from many other perspectives. The estimator is easily computable and applicable to such examples as extramarital data (high censoring, very large sample, many categorical regressors) or Stanford heart data (small sample, high censoring, categorical regressors). It does well in Monte-Carlo experiments and very sensibly in real-life examples. We believe this estimator will help procreate the presently scarce applications of the CQR models. Appendix B discusses other practical aspects of estimation and inference.

3.3. Large Sample Properties. The following assumptions are made in addition to equations (1), (5)-(7). For \( u_i(\tau) \equiv Y_i^* - X_i^j\beta(\tau) \) and all \( \tau, j \leq J \) of interest

Assumption 1. (i) \( \{(X_i, Y_i^*, C_i)\} \) are i.i.d; \( u_i(\tau) \) has density \( f_{u_i(\tau)}(u|X_i) \), which is bounded uniformly in \( X_i \), from above, away from zero, and continuous, uniformly in \( u \) near zero. \( u_i(\tau) \) has the unique \( r \)-th conditional quantile at \( 0 \). The support of the distribution of \( X_i \), \( X \), is compact. \( X_i \) includes a constant.

(ii) \( H_\eta(\tau) \equiv Ef_{u_i(\tau)}(0|X_i)X_iX_i'1[h(X_i) > (1 - \tau) + \eta] \) is positive definite, for a fixed constant \( \eta \in (0, \tau) \).
(iii) The pair of the binary model \( p \) and trimming constant \( c \) form a nontrivial envelope or a separating hyperplane of the propensity score: \( \exists c > 0, v \in (0, \tau), \)
\[
p(X_i' \gamma) > (1 - \tau) + c \text{ implies } h(X_i) > (1 - \tau) + v \text{ a.s.,}
\]
for any \( \gamma \) in a small neighborhood of \( \gamma_0 \equiv \lim_{\tau \to \gamma_0} \). \( E(X_i X_i' 1[p(X_i' \gamma_0) > (1 - \tau) + c] \) is invertible. \( p(\cdot) \) is strictly increasing and continuous. 

(iv) \( P(X_i' \alpha > v) \) is Lipshitz in \( \alpha \) uniformly in \( v \), for \( \alpha \) in an open neighborhood of \( \gamma_0 \) or \( \beta(\tau) \) and \( \hat{X}_i \) denoting \( \hat{X}_i \) and \( X_i \), respectively.

**Theorem 1.** Under the stated assumptions, as \( \delta_n \times \sqrt{n} \to \infty \) and \( \delta_n \downarrow 0 \)
\[
\sqrt{n} \left( \hat{\beta}_I (\tau) - \beta(\tau) \right) \overset{d}{\to} N \left( 0, H_0^{-1}(\tau) \Lambda_0(\tau) H_0^{-1}(\tau) \right)
\]
for finite \( I \geq 1 \), where \( \Lambda_0(\tau) \equiv \tau(1 - \tau)E(X_i X_i' 1[h(X_i) > 1 - \tau]) \). Furthermore, the same holds if any other consistent initial estimator \( \hat{\beta}_0(\tau) \) is used in step 2, provided sequence \( \delta_n \downarrow 0 \) and \( |\hat{\beta}_0(\tau) - \beta(\tau)| / \delta_n \overset{p}{\to} 0 \). Furthermore, the joint asymptotic distribution of several estimators for \( \tau_i, i \leq J \) is asymptotically normal, with covariance given by
\[
H_0^{-1}(\tau_i) \Lambda_0(\tau_i, \tau_j) H_0^{-1}(\tau_j),
\]
where \( \Lambda_0(\tau, \tau') \equiv [(\tau \wedge \tau') - \tau \tau'] E(X_i X_i' 1[h(X_i) > (1 - \tau) \vee (1 - \tau')]) \).

Thus the second part of the theorem allows for a wide range of alternative initial estimators \( \hat{\beta}_0(\tau) \).

**Remark 1.** Our proof uses an approach that is distinctly simpler than those in the cited literature and is therefore both short and straightforward to understand.

**Remark 2.** Assumptions (i) and (ii) are standard. Assumption (iii) allows the parametric "probability" model \( p(x' \gamma_0) \) to be moderately misspecified. The degree of misspecification is controlled by the constant \( c \). Assumption (iii) rationalizes the parametric first step, as we have discussed. It requires very weak smoothness conditions on the "probability model" \( p(\cdot) \): continuity and strict monotonicity.

**Remark 3.** No assumptions are made about the rate of convergence of \( \hat{\gamma} \) to its probability limit \( \gamma_0 \) or of \( \hat{\beta}_0(\tau) \) to \( \beta(\tau) \). The trimming device is designed to eliminate the bias and stochastic equicontinuity eliminates the impact of the variance of the preliminary steps. Assumption (iv) requires, for example, the distribution of \( \hat{X}_i' \alpha \) to respond smoothly to changes in \( \alpha \) in the vicinity of \( \gamma_0 \). The Lipshitz condition can be replaced by the weaker Holder continuity.

**Remark 4.** It is useful to take account of the structural risk sourced by the complexity of the envelope or separation models \( \varphi(x) = p(x' \gamma_n) - c \) that is increasing with \( n \), e.g. \( \hat{x}_n(x)' \gamma_n \) may be a power series or regression polynomial spline series. The result is preserved as long as \( x \mapsto p(\hat{x}_n(x)' \gamma_n) \) converges uniformly in \( C^r(X) \) to a fixed function, \( r > \dim(x)/2 \). The latter property may depend on particular estimation methods, many of which are discussed in Stone (1985), Cox (1988), Andrews and Whang (1990), Newey (1997), Chen and Shen (1998) and references therein. For brevity we do not reproduce the regularity conditions and particular methods of these references. Note that the resulting class of "envelope" models
\[
P \equiv \{ x \mapsto 1(\varphi(x) > c), \varphi \in \Theta, c \in [0, 1] \} \]
is Donsker as long as $\Theta$ is a compact set in $C^\infty(X)$ of boundedly differentiable functions of smoothness order $r = \dim(x)/2$ and $P(\phi(X) > c)$ is Lipshitz in $\phi$ w.r.t to the sup-norm in $\ell^\infty(X)$, uniformly in $c$. This is true since $L_2(P)$ bracketing number of $P$ is of the same order as that for $F = \{x \mapsto \phi(x) - c, \phi \in \Theta, c \in [0, 1]\}$ by monotonicity of the indicator function and Lipshitz property. The bracketing numbers for $F$ are given in example 19.9 in van der Vaart (1998) or Corollary 2.7.4. in van der Vaart and Wellner (1996). Hence if $r > \dim(x)/2$, the bracketing entropy integral for $P$ is finite and Donsker property holds in view of a constant envelope. Furthermore, Donsker property is preserved when $P$ is multiplied by a random variable with a constant envelope, as required in the proof.

3.4. Remarks on Robustness. The approach pursued here enjoys certain robustness properties, because it allows outliers and heavy tails for the dependent variable. The quantile regression estimator of Koenker and Bassett (1978), used in steps 2 and 3, is stable under arbitrary perturbations of $Y$ above and below the fitted line. This property is inherited from that of the ordinary sample quantiles and can be also advantageous when the distribution of $Y$ is heavy-tailed (e.g. in the extramarital example the Hill estimate of the tail index implies thick tails). In such cases least-square based methods, such as Buckley-James, may suffer a great deal. The first step is also robust in typical implementations. For example, as long as the censoring indicators $\delta_i$ are unaffected by perturbations in $Y_i$, the conditional MLE estimates are stable. Discriminant analysis will enjoy the same property. Even if the values of $\delta_i$ change, there have to be sufficiently many changes [$O(n)$] to distort the first step significantly.

3.5. Remarks on Computation. Computational expense of the estimator is comparable to linear least squares and appears to be of orders faster than computing other survival/censored regression estimators (based on our comparisons with survreg and coxph modules in S+). Step 1 is easily implemented by, for example, maximizing a convex, smooth quasi-likelihood. Steps 2 and higher all involve quantile regression with a very small computational expense due to Portnoy and Koenker (1997), whose algorithm is based on the interior point and preprocessing ideas. They show that both practical and theoretical computational times are of the same order as for linear least squares. QR software for R and S+ environments is available from statlib or http://www.econ.uiuc.edu/~roger/research/rq/rq.html. Other available software includes QR modules in STATA, SAS, Xplore.

3.6. Finite-Sample Properties and the Stanford Example. This subsection presents 1) a graphical bivariate simulation example, 2) a (standard) test of sensibility of our method on the well-known Stanford heart data-set.

3.6.1. A Simulation Example and Concordance Diagnostics. For the sake of clear visual illustration, we consider the case of median regression in three simple bivariate examples. These examples intend to clarify the workings of the method. They also help devise simple diagnostics. In each of the three models, there is a fixed censoring mechanism: $Y_i = Y_i^\star 1(Y_i^\star > 0)$. The data $(Y_i^\star, X_i), i \leq n = 200$ are generated as follows: $X_i = (1, \tilde{X}_i)'$, $\tilde{X}_i \sim N(1, 1)$, $u_i$ are i.i.d. and mutually independent, $u \sim N(0, 3)$, $\beta(\frac{1}{2}) = (0, 1)'$ and
• A. \( Y^* = X'\beta(\frac{1}{4}) + u, \)
• B. \( Y^* = X'\beta(\frac{1}{4}) + \tilde{X}u, \)
• C. \( Y_i^* = X'\beta(\frac{1}{4}) + u/\tilde{X}. \)

By construction, \( X'\beta(\frac{1}{4}) \) is the conditional median function. Model A is a standard normal model, while B and C are heteroscedastic normal models. 42%, 38% ,and 38% of the 200 observations are censored in the respective data sets, which isn't atypical for a lot of empirical applications.

Figure 2 illustrates our procedure. The rows correspond to models A-C. The first column plots the pre-censored data \((Y_i^*, X_i)\), and the solid line shows the true regression function \(x'\beta(\frac{1}{4})\). The data clouds are representative of the models’ nature.

A logistic probability model \( p(X_i'\gamma) \) is estimated by maximizing a quasi-likelihood function, using \( \delta_i \equiv 1 (Y_i > 0) \), \( \tilde{X}_i \equiv (X_i, X_i^2, X_i^3)' \). In the second column, "+" points represent \((Y_i, X_i)\) selected by the logistic envelope using \( p(X_i'\gamma) > 1 - \frac{1}{2} + c \). \( c \approx 0.1 \) is chosen such that about 80% of the observations with \( p(X_i'\gamma) > 1 - \frac{1}{2} \) are selected. The "o" points depict the rest of the sample. The second step regression quantile estimator, \( \hat{\beta}_0(\frac{1}{4}) \), is obtained by applying the Koenker-Bassett procedure to the selected sample. The resulting fit \( x'\hat{\beta}_0(\tau) \) is shown by the dashed line and contrasted with the solid line of the true regression function \( x'\beta(\frac{1}{4}) \).

The third column illustrates the third step. The "+" points now present \((Y_i, X_i)\) selected using \( X_i'\tilde{\beta}_0(\frac{1}{4}) > \delta_n \), where \( \delta_n \approx 0.7 \) is such that 90% of the observations with \( X_i'\tilde{\beta}_0(\frac{1}{4}) > 0 \) are selected. The "o" points represent the rest of observations. The third step estimator, \( \tilde{\beta}_1(\frac{1}{4}) \), uses the selected sample, and the fit \( x'\tilde{\beta}_1(\frac{1}{4}) \) is shown by the dashed line.

These figures are representative of the principle, and they seem to accord well with the asymptotic results described earlier. The logistic classifier, albeit an incorrect model for \( h(x) \), does well in separating out a good subset of observations - those with \( \tilde{X}_i > 0 \). The initial inefficient estimator \( \beta_0(\frac{1}{4}) \) is fairly reasonable. It serves as a classifier for the next step that separates a larger subset of good observations. Thus, the last step estimator uses a progressively larger sample and pools itself closer to the truth. This agrees well with the Monte-Carlo results discussed in the appendix.

Finally, the last column plots simple diagnostics that we found helpful. The column plots the logistic fit \( p(x'\gamma) \) vs. quantile fits \( x'\tilde{\beta}_0(\tau) - C_i \) (circles) and \( x'\tilde{\beta}_1(\tau) - C_i \) (triangles). The idea is to account for the concordance of different classifiers in choosing the trimming constants and deciding on the sensibility of the estimates. Each of the classifiers tries to separate out a set of observations for which \( X_i'\beta(\tau) > C_i \) or \( h(X_i) > 1 - \tau \). As a conservative classifier, the first step \( p(x'\gamma) \), with the addition of trimming, should be seen as one selecting a smaller set of observations. The subsequent classifiers should confirm all or almost all of this initial set.

The classifier concordance is presented by the placement of points in different quadrants of the plots A.4-C.4. For example, on the plot B.4, a large proportion of observations for which \( p(\tilde{X}_i'\gamma) > 1 - \tau \) is confirmed by the quantile classifiers, as seen in the upper-right quadrant. However, those points in the right-bottom corner, are dis-confirmed. Nonetheless, most of these are trimmed out by the trimming hurdle \( c \), \( p(\tilde{X}_i'\gamma) > 1 - \tau + c \). Hence the discordance
"from-the-right" is reduced or eliminated by such trimming. The left-bottom corners of A.4-B.4 represent the points agreeably disqualified by both the logistic and quantile classifiers.

The upper-left corners represent the discordance of the logistic and quantile classifier "from-the-left". In principle, this type of dissonance is not as pernicious as the one "from-the-right", since quantile classifier is meant to asymptotically separate a larger, better subset of observations. However, we find it prudent to reduce discordance "from-the-left" by adding the hurdle $\delta_n$. For example, on the plot A.4 the initial quantile classifier is overly optimistic in selecting a superset, rather than a subset of $\{i : X_i > 0\}$. The trimming factor $\delta_n$ reduces the discordance, and helps select a more appropriate set. In practice, since both the envelope and the quantile models are approximate models of reality, it appears prudent to reduce the discordance "from-the-left" as well.

Overall, if the concordance plots reveal a very drastic disagreement, it should serve as a sensible warning to revise the trimming constants, envelope models, or the models in question all at once.

3.6.2. The Stanford Example. The section considers a well-known Stanford heart transplant data set. The dataset is built into S+ (heart). Out of the 69 patients who received heart transplants, 45 had died by the closing date and are uncensored. Thus 35% of the observations are censored. The censoring times for the post-transplant survival are random observable. We shall contrast our results with the well-known and thorough study of Aitkin, Laird, and Francis (1983), and follow their definition of variables:

- (log) survival time, dependent variable: the log of the difference between the death and the transplant time. Censoring times are log differences between the last follow-up and the transplant dates.
- Age: Age of the patient at the time of acceptance into the program.
• Acc: Years since January 1, 1967 to acceptance into the program. This regressor may be seen as representing a technological progress.

• Surgery: indicator of a previous open heart surgery.

We estimate the CQR model:

\[ Q_{Y|X,C}(\tau) = (\alpha(\tau) + X'\theta(\tau)) \wedge C, \]

where \( C \) denotes the observable random censoring time. A benchmark model from survival analysis studied in Aitkin, Laird, and Francis (1983) is the accelerated failure time model:

\[ Q_{Y|X,C}(\tau) = (\alpha + \sigma F^{-1}(\tau) + X'\theta) \wedge C, \quad \forall \tau, \]

where \( F^{-1}(\tau) \) is the inverse of (a) the extreme value (Gumbel) or (b) the standard normal distribution function. Model (a) is thus a proportional hazard (PH) model with Weibull baseline hazard and model (b) – a non-PH AFT model. Among many models considered in Aitkin, Laird, and Francis (1983), these seemed to fit best. For comparisons, we shall use the estimates of \( \theta \) in model (14) reported in Aitkin, Laird, and Francis (1983), table 5. As we discussed, another way to robustly estimate \( \theta \) in model (14) is to take any \( \hat{\theta}(\tau) \), say median, or average over several quantile regression estimates \( \hat{\theta}(\tau) \) with different indices \( \tau \).

Because there are only few un-censored observations on patients with prior surgery, we could estimate the CQR model with the surgery regressor only up to the 50\%th percentile. To discuss higher quantiles, we also estimated the CQR model without the surgery regressor.

Figure 4 presents the results. The first four plots [read by rows] show the estimates of intercept coefficients, \( \hat{\alpha}(\tau) \), and of slope coefficients, \( \hat{\theta}(\tau) \), for \( \tau \) ranging from .1 to .5, for the model with age, acc and surgery as regressors. The shaded areas represent the pointwise 80\% confidence intervals. The bottom three figures plot the quantile coefficient estimates for the model with age and acc as regressors. The dotted lines present the Aitkin, Laird, and Francis (1983) estimates of quantile treatment/shift effects, \( \hat{\theta} \), in the model (14) with Weibull hazards (coefficients of PH model divided by the shape coefficient). The dashed lines are the estimates of quantile shift effect, \( \hat{\theta} \), in the normal version of (14). Note that the lines are horizontal since location-shift models impose constant treatment effects across quantiles.

Comparing the quantile shift estimates \( \hat{\theta}(\tau) \) in the CQR model with those in the models (14), \( \hat{\theta} \), across \( \tau \), we observe much qualitative and quantitative similarity. At the median, for example, the technological effects (acc) are both small and insignificant. The effects of age are both negative and significant, and the effects of previous surgery are both positive and significant. In other words, prior surgery and younger age seem to prolong the post-transplant survival. Overall, it is an encouraging fact that the 3-step CQR estimator produces results which compare well with a well-known, high-quality study.

Furthermore, as the CQR model (15) nests model (17) as a special case, we may confirm the general validity of the models considered in Aitkin, Laird, and Francis (1983), at least concerning the quantile treatment effects. In particular, the effects of prior surgery appear to be constant across all estimated quantiles. The effect of the time between January 1, 1967 to acceptance of the program is small and insignificant at all quantiles. However, we should point out that the age effects differ across quantiles. For low survival quantiles
the age treatment effects are positive, small, and statistically insignificant. For middle and higher survival quantiles the age effects are negative and significant. Yet this variation does not appear to (statistically) negate the location-shift models of Aitkin, Laird, and Francis (1983). Of course these findings warrant further careful examination of the age effects once more data becomes publicly available. Note that all of these quantitative and qualitative conclusions are well sustained when the surgery indicator is omitted from regression. See the bottom row in Figure 4.

Finally, the agreement of classifiers seems to be fairly high, as presented in the last two columns of the second row in figure 4. We plotted the logistic fit \( p(X_i'\beta) \) against the quantile fit \( C_i - X_i'\hat{\beta}_i (\tau) \). There are only a few discordant observations "from-the-right" or "from-the-left", almost all of which are eliminated by trimming.

![Figure 3](image)

**Figure 3.**

4. **Determinants of Extramarital Affairs: A CQR Analysis**

Extramarital affairs, an important social phenomenon, received much attention by anthropologists, psychologists, evolutionary biologists, sociologists, and economists. See Cronk (1991), DeLamater (1981), Fair (1978), Miller and Klein (1981), South and Lloyd (1995), Reiss, Anderson, and Sponaugle (1980) and many references therein. We present a retrospective analysis of the Redbook data-set on extramarital affairs. We shall mainly contrast our analysis with both the data and the model-analytic findings of Fair (1978). Fair (1978)
presents a utility-based optimization model of the time spent in the affair (affair intensity) as determined by preference for diversity, value of goods consumed in and outside marriage, labor and non-labor income, and time already spent with the spouse and paramour.

4.1. Data. The dataset has been collected by the Redbook magazine. Fair (1978) and DeLamater (1981) describe the collection procedures as well as the place of this data-set among only few similar data-sets. The data-set covers 6388 first-time married women, of which 68.5% reported to have had no extramarital affairs. This presents a very high degree of "censoring". We define all the variables as in Fair (1978) in order to facilitate comparisons with his statistical and model-analytic findings:

- Intensity of Affaire, dependent variable, defined as the number of different partners outside marriage times the approximate number of relationships with each partner, divided by the number of years in marriage. For 68.5% of respondents, it is equal to zero. For the rest of respondents, the density function is sketched by the histogram and a kernel estimator in figure 4.

![Affair intensity](image)

**Figure 4.**

Simple histograms of the following regressors are given in figure 5:

- Marriage Rating: respondents' rating of their marriage, on the scale from 1 to 5.
- Age, Years Married, No. of Children
- Religiousity: respondents' rating of their religiosity, on the scale from 1 to 4.
- Education: 9.0, 12.0, 14.0: grade school, high school, and some college; 16.0, 17.0, 20.0: college graduate, some graduate school, and advanced degree.

4.2. Models. The CQR model assumes the following form (with $C_i = 0$):

$$Q_{Y|X}(\tau) = (\alpha(\tau) + X'_{-1}\theta(\tau)) \vee 0. \quad (15)$$

That is, the conditional quantile function of the affair level is either zero or linear. This functional form is appealing, as we have discussed. We also consider a standard normal
\[ Q_{Y|X}(\tau) = (\alpha + \sigma \Phi^{-1}(\tau) + X_{-1}'\theta) \vee 0, \quad \forall \tau, \]  

(16)

where \( \Phi^{-1}(\tau) \) is the inverse of the standard normal distribution. Another benchmark model is the accelerated failure time model (for \( \exp(Y) \)) from survival analysis:

\[ Q_{Y|X}(\tau) = (\alpha + \sigma F^{-1}(\tau) + X_{-1}'\theta) \vee 0, \quad \forall \tau, \]  

(17)

where \( F \) is unspecified distribution function. It is easy to estimate the quantile shift effects \( \theta \) in this model by taking or averaging any of the estimates of \( \theta(\tau) \) in the CQR model. This will not be necessary, since neither this nor the normal model are supported by the data.

4.3. Estimation and Model Comparisons. To construct the initial envelope/classifier we examined the pairwise-plots of \( Y \) vs \( X \). Many of covariates appeared to be associated with higher dispersion of \( Y \), which lead us to consider a number of polynomial powers in the logistic model \( p(x'\gamma) \): \( \hat{x} \) consisted of \( x(i), x^2(i), x^3(i) \), and certain interactions \( x(i)x(j) \) that appeared to significantly improve the fit. Dimension of \( \hat{x} \) was 18, which is plausible in view of the large sample size. Sensitivity of the final estimates to further increasing the complexity of the envelope was negligible. The trimming constant \( c \) was set to \( c \approx .1 \) according to the rule described in the appendix. \( \gamma \) was estimated by conditional MLE.

Due to heavy censoring it was not possible to estimate all quantile coefficients \( \hat{\theta}(\tau) \). Identification depended on the nondegeneracy of the selected design matrix. This condition prevented considering quantiles lower than .5.

The results are summarized graphically in Figure 6. The solid line denotes the 3-step estimates of \( \hat{\beta}(\tau) = (\hat{\alpha}(\tau), \hat{\theta}(\tau))' \), \( \tau \in \{.4, ..., .9\} \), and the shaded region depicts the pointwise 90% confidence intervals. The dashed line presents the MLE estimates \( \hat{\theta} \) of the quantile shift effects in the normal model (16) obtained by Fair (1978).

\( \hat{\theta}(\tau) \) significantly vary across quantiles, especially at higher ones. This presents an evident violation of homoskedasticity assumption. Therefore, either model (16) or (17) are strongly
unsupportive of the data. It is still interesting to briefly comment on the behavior of ML estimates of the normal model. It is well known that the estimates are not robust to violations of both normality (e.g. heavy tails) and homoskedasticity – see Goldberger (1983), Hurd (1979) for proofs and simulation studies. In our example, we have both. It is interesting that for five out of eight variables, the estimates \( \hat{\theta}_j \) seem to correspond to fairly extreme quantile estimates \( \hat{\theta}_j(\tau) \), \( \tau \approx .9 \). For the marriage longevity variable, on the other hand, the estimate \( \hat{\theta}_j \) is far away from any of \( \hat{\theta}_j(\tau) \). Furthermore, in several cases sign of \( \hat{\theta}_j(\tau) \) changes across \( \tau \), so \( \hat{\theta}_j \), understandably, can not even match the direction of the quantile shift effect, since if the normal model (16) or location-shift model (17) were adequate, it would have been the case that \( \hat{\theta}(\tau) \approx \hat{\theta} \) for all \( \tau \). Thus, \( \hat{\theta}_j \) can hardly be given any meaning in the present setting. This finding is an empirical illustration to the earlier discussion on the breadth and flexibility of the CQR model.

Finally, the last row in Figure 6 presents the concordance plots. Overall, the concordance appears to be good. Only fairly small proportion of observations is discordant “from-the-left”, a major chunk of which is eliminated by the additional trimming. Discordance “from-the-right” is also mildly present, and part of it is also eliminated by trimming.

4.4. Analysis. The most exciting matter is the interpretation of the estimated quantile shift effects \( \hat{\theta}(\tau) \) (cf. Figure 6).

Religiosity effect is expectedly negative at all quantiles of affair intensity and especially strong at very high quantiles. As sociologists emphasized, the institutions and norms of Judeo-Christian doctrine have had a major influence on American families, endorsing a pro-creational and strictly-within-the-marriage orientation towards sexuality (e.g. DeLamater (1981)). This thesis is well supported by the data.

Education quantile shift effects, on the other hand, are more engaging. The effects are negative and strongly negative at high quantiles. Note that education and religiosity are weakly correlated (.14), but since we condition on religiosity, the education effects are net of this and other factors. Education effects are inexplicable within the Fair’s model, yet they appear to have a clear meaning in view of the relational (rather than recreational) perspectives towards a paramour among the more intelligent, educated, and not necessarily religious individuals (Reiss (1980), DeLamater (1981)).

The quantile treatment effects for age are nonpositive across all presented quantiles. Age effects are negative at the middle quantile and strongly negative at high quantiles. This means that the younger respondents are more likely to engage in an affair, especially in very intense ones, holding everything else fixed. In Fair’s analytic model, diversity considerations (“variety is a spice of life”) are incorporated in the utility but does not shed much light on the life-cycle considerations. On the other hand, the elaboration by DeLamater (1981) on the life-cycle perspective dynamics, within the social institutions and norms, conforms the present data-analytic findings.

Women with an occupation of higher socio-economic status are relatively more likely to engage in affairs, especially more intense ones. Explanations for this are to be looked at. One view is that such status creates an interactional advantage, increasing the hazard of an affair and subsequent marital dissolution (South and Lloyd 1995). Fair’s analytic model does
not necessarily yield predictions about the direction of the status effect (because he treats the status as proxy for labor income). To the extent that higher status is associated with non-labor income, or to the degree that income effects dominate substitution effects, Fair’s model may predict a positive effect.

*Husband’s occupational status* has very small positive or negligible effect across almost all quantiles, except at very extreme ones, where it becomes very negative (it is positive but insignificant at .95). Besides an (anecdotal) explanation that good men pick good wives,
women may value statusful husbands and pursue affairs, if at all, optimally to keep the risk of marital dissolution optimally low. On the other hand, Fair’s analytic model predicts the positive effects of a husband’s status (income) on the affair level, as a higher value of goods consumed in marriage causes wives to substitute labor activities for time spent with family and paramour. The Fair model, however, ignores the negative value of the dissolution option; which is real as other studies point out (South and Lloyd (1995)). Our quantile regression results suggest that Fair’s model explains only the middle quantiles, but does not apply to the high quantiles. It seems plausible that incorporation of the dissolution risk in Fair’s model, more along the lines of the Becker (1968) crime and punishment model, can make it conform to the present findings. Finally, it is also not implausible that more intelligent or statusful husbands can have better unobserved detection rates and the observed effect is the interaction of this hidden detection ability with the "cheating ability".

The effect of marriage longevity is slightly positive at .6-.8 quantiles and strongly negative at high quantiles. Fair, using behavioral considerations, postulates that marriage longevity may positively relate to diversity quest, leading to an increased affair level. However, it is not entirely clear why the effect is very negative at high quantiles. This may relate to the fact that only married and undivorced respondents were selected to the sample, so that the marriage longevity correlates with the marriage match quality and thus has a deleterious effect on affairs. Such an outcome would be a clear prediction of the search (for spousal alternatives) theory. Our finding thus partially disconfirms both the analytic and statistical predictions of Fair (1978) derived from the normal model.

5. Discussion

This paper had three goals. The first was to offer both theoretical and empirical perspectives on the global CQR models as a useful way to approximate the quantile treatment/shift effects in censored regression settings. The treatment variety is richer than the simple location-shift effects assumed by many commonly used models. In the CQR models, the covariates and the treatment can affect such features as scale, skewness, kurtosis, or, generally, the entire shape of conditional distribution.

The second goal was to offer an empirical CQR analysis of the determinants of a very serious social phenomenon – extramarital affairs. This is an important topic within sociology, psychology, and economics of marriage and family. To our regret, we found no previous implementable estimators that can be used in the settings of heavy censoring, many polytomous or continuous regressors, and large or small samples. Such data sets seem to prevail in many areas of applied statistics. This justified the pursuit of a practical, implementable, well-behaved estimator. The suggested estimator can be used to robustly estimate the global CQR models as well as many traditional models. Studying this estimator in large and finite samples, in the well-known Stanford heart transplant data-set, and developing the diagnostic tools, was the third goal.
Appendix

Below, $C$, $\text{const}$, and $K$ are generic positive constants. For notation sake, we set $C_i = 0$. General case is very similar.

**Appendix A. Proof of Theorem 1.**

**Part 1.** First consider $\hat{\beta}_0 (\tau)$. Rescaled statistic

$$Z_n^0 = \sqrt{n} (\hat{\beta}_0 (\tau) - \beta (\tau))$$

minimizes

$$Q_n (z, \hat{\gamma}) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_{in} (z) I [p(X'_i \hat{\gamma}) > 1 - \tau + c],$$

where

$$V_{in} (z) \equiv \sqrt{n} [p_r (\epsilon_i - X'_i z/\sqrt{n}) - p_r (\epsilon_i)] \text{ and } \epsilon_i \equiv Y_i - X'_i \beta (\tau) = \max (u_i, -X'_i \beta (\tau)).$$

The claim is, for finite $k$:

$$(Q_n (z_j, \hat{\gamma}), \ j \leq k) \overset{d}{\to} (Q_{\infty} (z_j), \ j \leq k),$$

where

$$Q_{\infty} (z) \equiv W' z + J z, \ W \overset{d}{=} N(0, \Lambda), \ J \equiv EF_\Lambda (0|X_1) X_1 X'_1 [p(X'_1 \gamma_0) > 1 - \tau + c], \ \Lambda \equiv \tau (1 - \tau) E X_1 X'_1 [p(X'_1 \gamma_0) > 1 - \tau + c].$$

By the convexity theorem (Knight (1999), Theorem 5, also see Pollard (1991)). But if $\hat{\gamma} = \gamma_0$, (18) follows by LLN, CLT, and some standard calculations, so it remains only to verify that

$$Q_n (z, \hat{\gamma}) - Q_n (z, \gamma_0) \overset{p}{\to} 0 \text{ for any fixed } z. \tag{19}$$

(a) For any fixed $z$, $\{Q_n (z, \gamma) - E Q_n (z, \gamma), \gamma \in \mathcal{G}\}$ is stochastically equicontinuous in $\gamma$, where $\mathcal{G} \equiv \{\gamma : |\gamma - \gamma_0| \leq \delta\}$ and $\delta > 0$ is small. Indeed, Type I functions (bounded variation functions of a single index, a VC subgraph class) in Andrews (1994) include the class $\mathcal{F} = \{x \mapsto 1[p(X'_1 \gamma) > 1 - \tau + c], \ \gamma \in \mathcal{G}\}$, hence by Theorem 2 of Andrews (1994) it satisfies Pollard's entropy condition with a constant envelope. This property is retained by the product of $\mathcal{F}$ with random variable $V_{in} (z)$, $V_{in} (z) \otimes \mathcal{F}$ by Theorem 3 in Andrews (1994),

since by assumption (i) $|V_{in} (z)|$ has a constant envelope:

$$|V_{in} (z)| \leq 2 |X_1 z| < \text{const} \tag{20}$$

Hence (a) is verified by Theorem 1 in Andrews (1994).

(b) Space $\mathcal{G}$ with pseudo-metric

$$p(\gamma_1, \gamma_2) \equiv \sup_{n \geq 1} \mathbb{E} \left| V_{in} (z) \times \left[ 1\{p(X'_1 \gamma_1) > 1 - \tau + c\} - 1\{p(X'_1 \gamma_2) > 1 - \tau + c\} \right] \right|^2$$

$$\leq C \times \left| P(X'_1 \gamma_1 > p^{-1}[1 - \tau + c]) - P(X'_1 \gamma_2 > p^{-1}[1 - \tau + c]) \right| + P(X'_1 \gamma_2 > p^{-1}[1 - \tau + c]) - P(X'_1 \gamma_1 > p^{-1}[1 - \tau + c]), \ X'_1 \gamma_2 > p^{-1}[1 - \tau + c])$$

$$\leq \text{const} \times \left\| \gamma_2 - \gamma_1 \right\|_2 \tag{21}$$

\*\*\*We note that Andrews allows for triangular sequences, which allows the random variable $V_{in} (z)$ to be indexed by $n$. Otherwise, we would have to refer to van der Vaart and Wellner (1996) 2.11.3 for dealing with functional classes indexed by $n$. Note that convexity allows us to treat $V_{in} (z)$ as random variables and not a function of $z$, in the sense that $z$ is fixed in the above arguments.
is totally bounded, where the first inequality follows from (20), and the second one follows from assumption (iv) by bounding each of the two terms, for example:

\[
\begin{align*}
&\left| P \left( X'_1 > p^{-1}[1 - \tau + c] \right) - P \left( X'_1 > p^{-1}[1 - \tau + c], X'_2 > p^{-1}[1 - \tau + c] \right) \right| \\
&\leq \left| P \left( X'_1 > p^{-1}[1 - \tau + c] \right) - P \left( X'_1 > p^{-1}[1 - \tau + c], X'_1 + X'_2 > p^{-1}[1 - \tau + c] \right) \right| \\
&\leq \left| P \left( X'_1 > p^{-1}[1 - \tau + c] \right) - P \left( X'_1 - K \times \| \gamma_2 - \gamma_1 \|_2 > p^{-1}[1 - \tau + c] \right) \right| \\
&\leq C \| \gamma_1 - \gamma_2 \|_2,
\end{align*}
\]

where \( |X'_1| (\gamma_1 - \gamma_2| \leq K \| \gamma_1 - \gamma_2 \|_2 \), since \( X_1 \) has a compact support \( X \).

(a) and (b) together implies that (e.g. Andrews (1994) equations 3.34 and 3.36):

\[
\sup_{|\gamma - \gamma_0| \to 0} \left| Q_n(z, \gamma) - Q_n(z, \gamma_0) - E Q_n(z, \gamma) + E Q_n(z, \gamma_0) \right| = o_p(1).
\]

Thus, to complete the proof of (19), it remains only to show that

\[
\left| E Q_n(z, \gamma) - E Q_n(z, \gamma_0) \right|_{\gamma = \tilde{\gamma}} = o_p(1).
\]  

We will show that for \( s_i(\gamma, \gamma_0) = 1 [ p(X'_1 \gamma) > 1 - \tau + c \] - 1 [ p(X'_1 \gamma_0) > 1 - \tau + c ]:

\[
E Q_n(z, \gamma) - E Q_n(z, \gamma_0) \equiv \sqrt{n} V_n(z) s_i(\gamma, \gamma_0)_{\gamma = \tilde{\gamma}} = O_p(\gamma - \gamma_0),
\]

Write \( \sqrt{n} V_n(z) \equiv -\sqrt{n} [ \tau - 1(\epsilon_i \leq 0) ] | X'_1 z | + \sqrt{n} [ - \eta(z) \{ X'_1 z - \epsilon_i / \sqrt{n} \} ] \equiv \sqrt{n} V'_n(z) + \sqrt{n} V''_n(z) \), where \( \eta(z) \equiv \{ 1(\epsilon_i \leq 0) - \epsilon_i \leq X'_1 z / \sqrt{n} \} \). Set \( \gamma \) close enough to \( \gamma_0 \). Then \( p(X'_1 \gamma) > (1 - \tau) + c \) implies \( h(X_i) > (1 - \tau) + v' \), and so does \( p(X'_1 \gamma_0) > (1 - \tau) + c \), by assumption (i) and (iii). Hence \( s_i(\gamma, \gamma_0) \neq 0 \) necessarily implies \( h(X_i) > (1 - \tau) + v' \), which implies \( X'_1 \beta(\tau) > v'' \), a.s. for \( v', v'' > 0 \) small, for all \( i \). Hence, as \( \gamma \) gets close to \( \gamma_0 \)

\[
E \left[ \sqrt{n} V'_n(z) s_i(\gamma, \gamma_0) | X_i \right] = E \left[ \sqrt{n} V''_n(z) 1( X'_1 \beta(\tau) > v'') | X_i \right] = s_i(\gamma, \gamma_0), \quad \text{uniformly in } i,
\]

since \( P(\epsilon_i \leq 0 | X_i, X'_1 \beta(\tau) > v'') = \tau \) [ if \( X'_1 \beta(\tau) > 0, \epsilon_i = \max(\epsilon_i, -X'_1 \beta(\tau)) \) has \( \tau \)-th conditional quantile at \( 0 \) ]. Also \( E \left[ \sqrt{n} V''_n(z) s_i(\gamma, \gamma_0) | X_i \right] \equiv \left[ E \left[ \sqrt{n} V''_n(z) 1( X'_1 \beta(\tau) > v'') | X_i \right] \times s_i(\gamma, \gamma_0), \quad \text{uniformly in } i \] [the second line follows by the standard calculations]. Therefore by assumptions (iii) and Lipshitz condition (iv):

\[
E E \left[ \sqrt{n} V''_n(z) s_i(\gamma, \gamma_0) | X_i \right] = O(E \left[ s_i(\gamma, \gamma_0) \right]) = O(\gamma - \gamma_0).
\]

(24) and (26) implies (22).

**Part 2.** It suffices to show the result for \( \hat{\beta}_1(\tau) \) with \( \hat{\beta}_0(\tau) \) as the selector. The proof for \( \hat{\beta}_1(\tau), I > 1 \) is identical. Proof for part 2 is similar to that of part 1, so only important differences are given. Notation undefined here is in part 1.

Rescaled statistic \( Z'_n = \sqrt{n}(\hat{\beta}_1(\tau) - \beta(\tau)) \) minimizes

\[
Q_n(z, \beta(\tau), \delta_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n V_n(z) 1[X'_i \beta(\tau) > \delta_n].
\]

Consider \( \delta_n \) as a parameter sequence. Proceed identically as in part 1 up to equation (19), only replacing \( 1[p(X'_1 \gamma) > 1 - \tau - c] \] by \( 1[X'_1 \beta(\tau) > \delta_n] \), \( 1[p(X'_1 \gamma) > 1 - \tau - c] \) by \( 1[X'_1 \beta(\tau) > 0] \) and \( W, J, \Lambda \) by \( W \equiv N(0, H_0^{-1} \Lambda_0 H_0^{-1}) \), \( H_0, \Lambda_0 \). It remains to show

\[
Q_n(z, \beta(\tau), \delta_n) - Q_n(z, \beta(\tau), 0) \overset{p}{\to} 0, \text{ for any fixed } z.
\]

(a) For any fixed \( z \), \( Q_n(z, \beta, \delta) = E Q_n(z, \beta, \delta), (\beta, \delta) \in B \times D \) is stochastically equicontinuous in \( \beta, \delta \), where \( B \equiv \{ \beta : |\beta - \beta(\tau)| \leq C' \} \), \( D \equiv \{ \delta : 0 \leq \delta \leq C'' \} \) and \( C', C'' > 0 \) are small. Indeed, Type I functions in Andrews (1994) (VC subgraph classes) contain \( F = \{ x \mapsto 1[x' \beta > \delta], (\beta, \delta) \in B \times D \} \). Hence \( F \) has
a finite uniform covering entropy integral and a constant envelope, i.e. satisfies Pollard's entropy condition. This property is retained by the product of space $\mathcal{F}$ with random variable $V_n(z)$: $V_n(z) \otimes \mathcal{F}$, by Theorem 3 in Andrews (1994), since $|V_n(z)|$ has a constant envelope. (a) is verified by Theorem 1 in Andrews (1994).

(b) Space $B \times D$ is totally bounded under the $L_2$ pseudometric:

$$
\rho((\beta_1, \delta_1), (\beta_2, \delta_2)) \equiv \sup_{\pi} \left| E[V_n(z) x [1 \{X'_1 \beta_1 > \delta_1\} - 1 \{X'_1 \beta_2 > \delta_2\}]^2]\right|
\leq \text{const} \times \|\beta_2 - \beta_1\|_2 + \text{const} \times \|\delta_2 - \delta_1\|_2,
$$

where (28) follows from (20) and from assumption (iv), analogously to the proof of (21), treating $\delta$ as shifting the intercept parameter.

(b), along with (a), implies (e.g. Andrews (1994) eq (3.34) and (3.36)):

$$
\sup_{|\beta - \beta(\tau)| \to 0} \left| Q_n(z, \beta, \delta_n) - Q_n(z, \beta(\tau), 0) - E Q_n(z, \beta, \delta_n) + E Q_n(z, \beta(\tau), 0)\right| = o_p(1).
$$

Thus, to complete the proof of (27), it remains to show that

$$
E Q_n(z, \beta, \delta_n) - E Q_n(z, \beta(\tau), 0) = o_p(1) \quad (29)
$$

Let $s_i(\beta, \beta(\tau)) \equiv 1(X'_i \beta > \delta_n) - 1(X'_i \beta(\tau) > 0)$. By assumption on the sequence $\delta_n$, w.p. $\to 1$, $\hat{\beta}_n(\tau)$ is inside the ball with radius $\kappa \delta_n$, centered at $\beta(\tau)$, where $\kappa > 0$ is small. By the compactness assumption on $X_i$, $\kappa'$ can be chosen so that w.p. $\to 1$, $\sup_{\beta \in \mathcal{X}} |X'_i(\hat{\beta}_n(\tau) - \beta(\tau))| < \frac{1}{2} \delta_n$. Set $\beta$ inside this ball. Then for small enough $\kappa'$ chosen as such, $x'_i \beta > \delta_n$ implies $x'_i \beta(\tau) > 0$, and $x'_i \beta(\tau) \leq 0$ implies $x'_i \beta \leq \delta_n$. Thus $s_i(\beta, \beta(\tau)) \neq 0$ necessarily implies $X'_i \beta(\tau) > 0$ a.s.. Hence, uniformly in $i$,

$$
E\left[\sqrt{n} V''_{\tau}(z) s_i(\beta, \beta(\tau)) | X_i\right] = E\left[\sqrt{n} V''_{\tau}(z) 1(X'_i \beta(\tau) > 0) | X_i\right] \times s_i(\beta, \beta(\tau)) = 0,
$$

since $P[c_i \leq 0 | X_i, X_i(\beta(\tau)) > 0] = 0$. Also $E\left[\sqrt{n} V''_{\tau}(z) s_i(\beta, \beta(\tau)) | X_i\right] = E\left[\sqrt{n} V''_{\tau}(z) 1(X'_i \beta(\tau) > 0) | X_i\right] \times s_i(\beta, \beta(\tau)) = O\left[|f_n(0 | X_i) z'_i X_i z 1(X'_i \beta(\tau) > 0)\right] \times s_i(\beta, \beta(\tau)) = 0,
$$

unifomly in $i$ [the last equality again follows by the standard calculations]. Therefore by assumptions (iii) and Lipshitz condition (iv)

$$
E E\left[\sqrt{n} V''_{\tau}(z) s_i(\beta, \beta(\tau)) | X_i\right] = O\left[E s_i(\beta, \beta(\tau))\right] = O(\beta - \beta(\tau)) + O(\delta_n).
$$

(30) and (31) give (29).

**Part 3.** Finally, to show joint convergence of $Z_n(\tau_j) = \sqrt{n}(\hat{\beta}_n(\tau_j) - \beta(\tau_j))$, $i \leq J$ for either step, proceed as before by defining variable $Z_n = (Z_n(\tau_i), i \leq J)$ as minimizing the objective function: $Q_n(z) = \sum_{i \leq J} \bar{Q}_n(zi)$ over $\tau = (zi, i \leq J)$, where $\bar{Q}_n$ are objective functions defined as in part 1 or part 2. Repeat the arguments in parts 1 and 2, noting that sum of s.e. processes is s.e. (see e.g. Knight (1999)).

**Appendix B. Practical Details of Estimation and Inference**

Inference: Because of distributional equivalence with the Powell CQR estimator, all of the inference procedures developed by Powell apply without modifications. The bootstrap inference of Bliias, Chen, and Ying (2000) can be extremely useful in practice. All inference procedures are as for the standard quantile regression procedure with the only difference being that the selected (rather than complete) sample is used. Therefore, a user of the standard QR software need not make any modification – the standard errors, confidence intervals produced by the last step QR routine are all valid.

Model $p(\cdot)$ and Trimming Constant: Parametric model $p$ in step 1 should be reasonably good. Goodness of fit checks are supplied in all the standard statistical packages and are easy to carry out. It is always possible to achieve a reasonably good fit using a series approximation of probability model. The theory requires $c$ to be not very small, although the Monte-Carlo work shows that $c = 0$ is also quite sensible. In practice, one should check sensitivity of the estimates $\hat{\beta}_0(\tau)$ to increasing the constant $c$. Also, $c$ shouldn’t be too big, for we may
select very small $J_0$. The better the parametric fit is, the smaller $c$ can be set, improving efficiency of the initial \( \hat{\beta}_0 \). A sensible rule for choosing $c$ is to compare the size of the selected sample $J(c) = \{ i : p(X_i' \gamma) > 1 - \tau + c \}$ for $c = 0$ and other values. Choosing $c = q$-th quantile of all $p(X_i' \gamma)$ s.t. $> 1 - \tau$, appears to be sound, as it gives a control of percentage of observations from $J(0)$ can be thrown out: $\#J(c)/\#J(0) = (1 - q)\%$. This rule, with $q = 10\%$, was employed in simulations. We also set \( \delta_n \) corresponding to $q = 5\%$.

**APPENDIX C. MONTE-CARLO EXPERIMENTS**

Finally, table I reports the result of a small monte carlo experiments.\(^{10}\) The model we consider was a standard location-scale model with an error term hit by a linear-quadratic heteroscedastic scale, for $X_i = (1, \tilde{X}_i)': Y_i = X_i' \beta + \epsilon_i$. We draw $\tilde{X}_i \in R^3$ from independent standard normal distributions, truncated by $\{X_i : ||X_i|| < 2\}$. The error term has the multiplicative heteroscedasticity structure: for $u_i \sim N(0, 25)$, $\epsilon_i = u_i \times \left( 1 + 0.5 \sum_{j=1}^3 \left( \tilde{X}_i + \tilde{X}_j^2 \right) \right)$. The true parameter vector is chosen at $(1, 1.5, -1, -0.5, 25)$, and the censoring point is 0.75. We used $X$ and $X^2$ in the parametric propensity score regressions. We experimented with different probability models including logit, probit and linear models. However, the type of probability model used has very little effect on the performance of the estimators. Therefore, in table I we only report the results for the probit selection model. Note that due to the heteroscedasticity error structure, even the probit model is not consistent with the true propensity score. We report the initial step, the first step and the third step estimators, and compare them to the Buchinsky and Hahn (1998) estimator and the Powell estimator.

Notably, the results from the monte carlo simulation show that for sample size 100, the step-3 (I=1) estimator outperforms all other estimators in terms of root mean square errors(RMSE). Its mean absolute deviation(MAE), both mean and median biases, are comparable to other estimators. Iterating to step 5(I=5) increases both root mean square error and mean absolute deviation, at the benefit of reducing both mean and median biases. Overall the step 5 estimator still compares favorably to both the Buchinsky and Hahn (1998) estimator and the Powell estimator. The initial inefficient estimator (I=0) does fairly poorly, as expected. In a large sample, $n = 400$, the 3 estimator and the Buchinsky and Hahn (1998) estimator perform equally well and are both more favorable than other estimators in all dimensions. To conclude, the 3-step estimator does better in a small sample, and very well in a large sample.

**References**


---

\(^{10}\)The set up of our simulation is similar to those of Buchinsky and Hahn (1998).


### Table I

**Monte Carlo Simulation Results with Five Regressors for .50 Quantile (1,000 Repetitions.)**

*Normal Heteroskedastic Errors*

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Slope</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I=0</td>
<td>I=1</td>
<td>I=3</td>
<td>BH *</td>
<td>I=0</td>
<td>I=1</td>
<td>I=3</td>
</tr>
<tr>
<td>n=100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>5.08</td>
<td>2.15</td>
<td>2.79</td>
<td>3.25</td>
<td>3.93</td>
<td>2.69</td>
<td>1.89</td>
</tr>
<tr>
<td>Mean bias</td>
<td>4.16</td>
<td>1.46</td>
<td>1.59</td>
<td>1.38</td>
<td>0.7</td>
<td>0.2</td>
<td>-0.08</td>
</tr>
<tr>
<td>MAE</td>
<td>4.23</td>
<td>1.7</td>
<td>1.86</td>
<td>1.89</td>
<td>1.55</td>
<td>2.02</td>
<td>1.39</td>
</tr>
<tr>
<td>Median bias</td>
<td>3.71</td>
<td>1.43</td>
<td>1.33</td>
<td>1.54</td>
<td>1.22</td>
<td>0.03</td>
<td>-0.43</td>
</tr>
<tr>
<td>n=400</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>2.28</td>
<td>1.05</td>
<td>1.21</td>
<td>1.28</td>
<td>1.31</td>
<td>1.13</td>
<td>0.88</td>
</tr>
<tr>
<td>Mean bias</td>
<td>1.94</td>
<td>0.64</td>
<td>0.75</td>
<td>0.62</td>
<td>0.74</td>
<td>-0.18</td>
<td>-0.28</td>
</tr>
<tr>
<td>MAE</td>
<td>1.99</td>
<td>0.82</td>
<td>0.88</td>
<td>0.8</td>
<td>0.81</td>
<td>0.89</td>
<td>0.69</td>
</tr>
<tr>
<td>Median bias</td>
<td>1.83</td>
<td>0.66</td>
<td>0.71</td>
<td>0.64</td>
<td>0.78</td>
<td>-0.19</td>
<td>-0.43</td>
</tr>
</tbody>
</table>

Note: Sample sizes are denoted by n(100, 400). The last two columns are from Buchinsky and Hahn(1998) Table 2. BH denotes their estimator Cva, which uses a cross-validated bandwidth adjusted to the under-smoothing assumption. Powell* (ILPA*) is the Powell estimator, or more precisely, an estimator obtained by iterated linear programming.