Exploring Transition Data through Quantile Regression Methods: an Application to U.S. Unemployment Duration *

José A.F. Machado
Universidade NOVA de Lisboa, Faculdade de Economia †
and
Pedro Portugal
Banco de Portugal
Universidade NOVA de Lisboa, Faculdade de Economia

April, 2002

Abstract

Quantile regression constitutes a natural and flexible framework for the analysis of duration data in general and unemployment duration in particular. Comparison of the quantile regressions for lower and upper tails of the duration distribution affords important insights into the different determinants of short or long-term unemployment. Using quantile regression techniques, we estimate conditional quantile functions of U.S. unemployment duration; then, resampling the estimated conditional quantile process, we are able to infer the implied hazard functions. The proposed methodology proves to be resilient to several misspecifications that typically afflict proportional hazard models, such as neglected heterogeneity and baseline misspecification. Overall, the results provide clear indications of the interest of quantile regression to the analysis of duration data.

KEYWORDS: Quantile Regression, Duration Analysis, Unemployment Duration

JEL CODES: C14, C21, C41, J64

*We would like to thank seminar participants at the Universidade Católica Portuguesa, Universidade do Porto, Instituto Superior de Economia e Gestão and NOVA, and participants at the ASSET Meeting (Lisbon, 2000), European Meeting of Statisticians (Madeira, 2001) and the European Meeting of the Econometric Society (Lausanne, 2001). Comments by Joshua Angrist, Victor Chernozhukov, Marcelo Fernandes, Miguel Gouveia, Roger Koenker, José Mata, and João Santos Silva were greatly appreciated. Machado is consultant for the Research Department of the Bank of Portugal. The authors also gratefully acknowledge the partial financial support from the Fundação para a Ciência e a Tecnologia. The usual disclaimer applies.

†Corresponding author: José A.F. Machado, Faculdade de Economia, Universidade NOVA de Lisboa, Campus de Campolide, 1099-032 Lisboa, Portugal. E-mail: jafm@fe.unl.pt

1 Introduction

The number of applications of quantile regression techniques has greatly increased in recent years (for an overview of the range of those applications see, for instance Fitzenberger et al. (2001)). Labor economics has been one of the most popular fields for applications, but attention has been almost exclusively devoted to the study of wage equations (see, for example and with no claim to being exhaustive, Buchinsky (1994) and (2001), Chamberlain (1994), Fitzenberger and Kurz (1997) and Machado and Mata (2001)).

The chief aim of this paper is to explore the potential of models for conditional quantile functions, or quantile regression models, as a tool for analyzing duration data. Quantiles seem quite appropriate to this analysis for two major reasons. First, as the seminal work of Powell (1984, 1986) reveals, quantile regression is particularly well equipped to perform consistent inferences with censored data, a typical situation in duration studies. Second, the methodology estimates the whole quantile process of duration time conditional on the attributes of interest. Quantile functions constitute, as does the more traditional hazard function, a complete characterization of the (conditional) distribution of duration time or, if one wishes, of the survivor function.

Estimation of models for conditional quantiles provides, therefore, a characterization of the entire conditional distribution of duration time. This characterization is robust, for it does not rest on strong distributional assumptions and, yet, parsimonious, for it is semiparametric in nature. More importantly, it is quite flexible, as it allows the effects of the covariates to differ at different points of the distribution and, thus, it may capture "transient effects" that is, instances where a covariate exert a significant influence on one tail of the distribution but not on the other. This feature makes conditional quantiles a natural statistical framework for analyzing how short-term unemployment—that is, duration in the left tail of the distribution of unemployment times—differs from long-term unemployment—that is, the right tail of the distribution.

In addition to these potential informational gains, quantile regression methods offer three features of great interest for duration analysis. First, it is possible to cast the results from quantile regressions in the typical frames of other modeling approaches; in particular, it is possible to estimate the survivor function, residual duration and the hazard function implied by a given model for conditional quantiles. Second, the estimators of the covariates' effects on duration are robust to neglected heterogeneity uncorrelated with the covariates. Finally, when applied to Proportional hazard models, the quantile regression estimators are resilient to misspecification of the baseline hazard.

The works by Horowitz and Neumann (1987, 1989) constitute early attempts at using quantile estimates for employment duration. However, somehow, they do not appear to have made their way into the mainstream econometric analysis of duration. Be that as it may, the emphasis there was the consistent estimation of a parameter vector in the presence of censoring, rather than the exploitation of the full potential of quantile regression as a tool for the statistical analysis of conditional distributions. Two recent papers by Koenker and Geling (2001) and by Koenker and Billias (2001), more in the vein of the present one, bear witness to the renewed interest in the topic. Apart from the issues of functional form misspecification and unobserved individual heterogeneity that we address in this study, the main differences between this paper and the studies by Koenker and Geling (2001) and by Koenker and Billias (2001) are three-fold: first, we tackle the problem of censored duration (incomplete durations are at the center stage of survival analysis); second, we offer an alternative method for estimating the hazard and survival functions; and third, we present the quantile regression estimates in a framework that makes it comparable to the proportional hazard model.

The paper is organized as follows. The next section provides the main ideas of quantile regression and discusses the estimation of the hazard function implied by a regression quantile model. There, we also address the issue of the impact of neglected heterogeneity on the quantile regression estimators and present a small simulation study of the resiliency of quantile regression estimators of Proportional Hazard models to the baseline misspecification. In Section 3 we illustrate the approach with a well known and important data set—the U.S. "Displaced Workers Survey"—in order to highlight the potential information gains from using quantile regression in duration analysis. Special attention is devoted to expressing the output of quantile regressions in terms of survivor and hazard functions. Section 4 concludes.

2 Quantile Regression Models in Duration Data Analysis

2.1 Regressions

Models specified in terms of hazard functions undoubtedly dominate the analysis of duration data. Yet, in some instances, regression-type models may prove natural and useful. Regression models for the duration time are typically framed in a strict parametric setting. Let T be the duration of stay in a given state, and x_i ($x_{1i} \equiv 1$) be the vector of covariates for the ith observation. A parametric regression model assumes that

$$z(T_i) = x_i'\beta + \sigma\epsilon_i \tag{1}$$

where, β and σ are unknown parameters, $z(\cdot)$ is a transformation function and ϵ is a zero mean and unit variance random variable with density f, not depending on x, (e.g., Gaussian, lognormal, smallest extreme value, Weibull or exponential). A leading example of this class is the Accelerated Failure–Time (AFT) model where

$$\log T_i = x_i'\beta + \sigma\epsilon_i \tag{2}$$

and f is left unspecified. The Proportional hazard (PH) model with Weibull baseline also fits in the class, as it is equivalent to the Accelerated Life model with ϵ being the log of a unit Exponential variate.

The set-up above is restrictive in two main ways. First, it assumes a known duration distribution f so that the model may be estimated by maximum likelihood. As is well known, the resulting estimators are "optimal" if the model is correctly specified but lack robustness to departures from the assumed distribution.

Second, and perhaps even more importantly, (1) assumes that only the conditional mean of $z(T)$ depends on the covariates. In technical terms, the distribution of the duration time conditional on the covariates is restricted to the translation family that is, all the heterogeneity in the distribution of duration time for different levels of the covariates is assumed to be captured by mere location shifts (Manski, 1988). To put it plainly, the distributions corresponding to different individuals differ only on location; other distributional attributes such as scale, skewness or tail behavior are deemed independent of the conditioning variables.

Quantile regression (QR) directly addresses these two limitations of a strict parametric approach. For $p \in (0, 1)$, let

$$Q_T(p|x) = \inf\{t|F_{T|X}(t|X = x) \geq p\}$$

denote the pth quantile of the conditional distribution of T given $X = x$ ($F_{T|X}(t|X = x)$). We consider statistical models specifying

$$Q_T(p|x) = g(x'\beta(p)) \quad (3)$$

where $g(\cdot)$ is a monotone link function, known possibly up to a finite number of parameters $\lambda(p)$, and $\beta(p)$ is a vector of QR parameters, varying from quantile to quantile. Owing to the equivariance property of quantiles to monotone transformations, if we denote $y(\cdot) \equiv g^{-1}(\cdot)$, (3) may be written as

$$Q_{y(T)}(p|x) = x'\beta(p) \quad (4)$$

or, in a regression-type format,

$$y(T) = x'\beta(p) + u(p) \quad (5)$$

where $u(p)$ is an error term with $Q_{u(p)}(p|x) = 0$.

As an illustration, it is clear that the AFT model (2) implies that

$$Q_{\log T}(p|x) = x'\beta + \sigma Q_\epsilon(p),$$

thus revealing that all the conditional quantiles of $\log T$ are parallel, in that the difference between any two does not depend on x.

By contrast, the QR model (3), allows for the covariates to have different effects on different points of the distribution. To illustrate this general point, consider a simple treatment effects problem,

$$Q_{\log T}(p|d) = \alpha(p) + \delta(p)d$$

with $d = 1$ for treatment and $d = 0$ for the control group. The QR framework is flexible enough to allow for, say, $\delta(0.25) > 0$ but $\delta(0.9) = 0$—the treatment being effective on left tail but not on the right tail of the duration distribution. It is interesting to note that this example describes a kind of situation that may be modeled by a hazard model with "transient effects" (see *e.g.*, Cox and Oakes, (1984)).

However, the QR model ought not to be taken as a mere heteroscedastic extension of the AFT model. It is true that (3) allows for the scale of the conditional duration distribution to change with x; but it also allows for the skewness, tail behavior and, in general, the whole distribution to depend on the covariates.

In addition to these potential informational gains, quantile regression methods offer three features of great interest for duration analysis:

- They provide ways of estimating typical outputs from other modeling approaches, namely the hazard function.
- The estimators of the covariates' effects on duration are robust to neglected heterogeneity uncorrelated with the covariates.
- In the context of a PH model, the estimators are resilent to misspecification of the baseline hazard.

The next sections explore in detail each of these points.

2.2 Estimation of the Hazard Function

Model (3) provides a complete characterization of the (conditional) distribution of duration time T or, if one wishes, of the survivor function, (obviously, $Q_T(p|x)$ is the $(1-p)$th quantile of the conditional survivor function). The hazard function,

$$h(t|x) = \frac{f_{T|X}(t|x)}{1 - F_{T|X}(t|x)}$$

provides still another characterization of the same probability distribution. Since it constitutes the most popular frame for duration analysis, it is important to relate it to models for the conditional quantile function (CQF).

Making $t = Q_T(p|x)$, one has,

$$h(p|x) \equiv h(Q_T(p|x)|x) = \frac{f_{T|X}(Q_T(p|x)|x)}{1-p}. \tag{6}$$

Total differentiation of the identity $F_{T|X}(Q_T(p|x)|X = x) = p$ and substitution in (6) yields,

$$h(p|x) = \frac{1}{(1-p)\partial Q_T(p|x)/\partial p}. \tag{7}$$

Equation (7) provides the basic relationship between conditional quantile functions and hazard functions, and suggests a way of estimating the hazard function implied by a given QR model.

The simplest approach is based on a proposal by Siddiqui (1960) to estimate the sparsity function (the inverse of the density function). Using Siddiqui's estimator of the sparsity function —the inverse of the density function—(Siddiqui, 1960)we have,

$$\hat{h}(p|x) = \frac{2\nu_n}{(1-p)[\hat{Q}_T(p+\nu_n|x) - \hat{Q}_T(p-\nu_n|x)]} \tag{8}$$

where ν_n is a bandwidth depending on the sample size n (which tends to 0 as $n \to \infty$) and $\hat{Q}_T(\cdot|x)$ denotes the estimated QR, (see, also Hendricks and Koenker, 1992 and Koenker and Geling, 2001).

Alternatively, one may resort, as we do in this paper, to the resampling procedures proposed in Machado and Mata (2000). The proposed estimator is based on a simulated random sample, $\{T_i^\star, i = 1, \ldots, m\}$, from a conditional distribution of duration time that is consistent with the restrictions imposed on the conditional quantiles by the QR model. To this sample the usual methods of density estimation and hazard function estimation may be applied (see, e.g, Silverman 1986). In detail, the procedure is as follows:

1. Generate m random draws from a Uniform distribution on $(0, 1)$, $\pi_i,\ i = 1, \ldots, m$;
2. For each π_i estimate the QR model (4), thereby obtaining m vectors $\hat{\beta}(\pi_i)$;
3. For a given value of the covariates, x_0, the desired sample is,

   $$T_i^\star \equiv \hat{Q}_T(\pi_i|x_0) = g(x_0'\hat{\beta}(\pi_i))$$

   for $i = 1, \ldots, m$.

The theoretical underpinnings of this procedure are quite simple. On the one hand, the probability integral transformation theorem from elementary statistics implies that one is simulating a sample from the (estimated) conditional distribution of T given $X = x_0$. On the other hand, the results in Bassett and Koenker (1986) establish that under regularity conditions the estimated conditional quantile function is a strongly consistent estimator of the population quantile function, uniformly in p on a compact interval in $(0, 1)$.

Once the sample $\{T_i^\star, i = 1, \ldots, m\}$ was generated, the hazard function may be estimated as (Silverman, 1986, p.148),

$$\hat{h}(t|x) = \frac{f^\star(t)}{1 - F^\star(t)}$$

where $f^\star(t)$ is the usual kernel density smoother of $T_i^\star$,

$$f^\star(t) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{t - T_i^\star}{h}\right)$$

and the distribution function estimator is,

$$F^\star(t) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{K}(\frac{t-T_i^\star}{h})$$

with

$$\mathcal{K}(u) = \int_0^u K(v)dv.$$

In situations where, due to censoring, the top or bottom quantiles cannot be consistently estimated, the π_i in step 1 must only be generated in the relevant range and the estimated function must be adequately rescaled.

Besides hazard functions, other standard outputs of duration analysis such as survivor function, residual duration and mean duration are also quite easily estimated from a quantile model such as (4). For instance, given an estimate of the quantile function of T, $\hat{Q}_T(p|x)$, the quantile process of the survivor time conditional on x can be estimated by $\hat{Q}_T(1-p|x)$ which, upon "inversion", yields an estimate of the the survivor function (see, Bassett and Koenker, 1986). The mean duration conditional on x can be estimated as $\int_0^1 \hat{Q}_T(p|x)dp$ which can be easily computed by Monte-Carlo methods; by the same token, taking the simple treatment effects model presented above as an example, $\int_0^1 \delta(p)dp$ represents the effect of the treatment on the mean log-duration. Likewise, the distribution of the residual duration—i.e., the duration of all those that have survived longer than $Q_T(p^*|x)$, for a given p^*—may be summarized by $\int_{p^*}^1 Q_T(p|x)dp$.

2.3 Neglected heterogeneity

The PH model with unobserved heterogeneity may be written as,

$$h(t|x) = h_0(t)\exp(-(\alpha + x'\beta + u))$$

where $h_0(t)$ is the baseline hazard function and u is an unobserved random variable independent of the covariates (x).

This model is equivalent to

$$\log H_0(t) = \alpha + x'\beta + u + \epsilon$$

where ϵ is a Type-I extreme value variate, and

$$H_0(t) = \int_0^t h_0(s)ds$$

which equals t when the baseline is Weibull, (see, e.g. Lancaster(1990)).

Clearly,

$$Q_{logH_0(t)}(p|x) = (\alpha + Q_u(p) + Q_\epsilon(p)) + x'\beta \tag{9}$$

and thus only the intercept is affected by the presence of unobserved heterogeneity. This result is obviously not specific to QR estimators. Indeed, as long as the baseline hazard is correctly specified, it holds for any estimator that does not rely on specific distributional assumptions about the error terms. For instance, as long as there is no censoring, it also holds for the least squares estimator.

The crucial qualifier is the correct specification of $h_0(t)$. Ridder (1987), analyzes this interplay between unobserved heterogeneity and specification error. Next, by means of a small simulation study, we evaluate the resiliency of QR estimators to mispecification of the baseline function.

2.4 Baseline misspecification

In this section we analyze the consequences of misspecifications of transformations of the dependent variable in models such as (9). Specifically, we consider the following experiment. The data was generated by

$$B(Y;\lambda) = a + bx + \epsilon \tag{10}$$

with x and ϵ standard normal variates. $B(Y;\lambda)$ is the Box-Cox transformation of Y. Using the equivariance property of quantiles, the implied population quantile function of Y is

$$Q_Y(p|x) = B^{-1}(a + bx + Q_\epsilon(p);\lambda) \tag{11}$$

with $B^{-1}(z;\lambda)$ denoting the inverse Box-Cox transformation

$$B^{-1}(z;\lambda) = \begin{cases} (1+\lambda z)^{(1/\lambda)} & \lambda \neq 0 \\ \exp(z) & \lambda = 0 \end{cases}.$$

The coefficients of the pth quantile regression of Y on x provide estimates of the slope of the pth population quantile of Y given $x = \bar{x}$, the sample average of the covariates. Table 1 shows these slopes for different values of transformation parameter, (λ =2, 1/2, 0, 1 and -1), that is

$$\partial Q_Y(p|x)/\partial x = \begin{cases} [1+\lambda(a+Q_\epsilon(p))]^{(1/\lambda)-1} & \lambda \neq 0 \\ \exp(Q_\epsilon(p)) & \lambda = 0. \end{cases} \tag{12}$$

The parameters were set to $a = 10$ and $b = 1$, except that for $\lambda = 0$ where $a = 0$. Of course for $\lambda = 1$ the slope is also 1 at every quantile.

The experiment purports to compare the QR estimators of that slope with the maximum likelihood estimators (ML). The QR estimators were obtained regressing Y on an intercept and x for different values of $p \in (0,1)$ and for the different data sets. The ML estimators of the slopes were obtained by plugging-into (11) the MLE of λ, a and b in (10) and then taking derivatives as in (12).

The comparison was based on the median estimates and on its inter-quartile range in 1000 replications. Table 2 shows the results. Both methods perform well for all the data generating process, with the exception of that corresponding to $\lambda = -1$. The interesting feature, however, is that even though it requires much less prior knowledge, the QR estimator clearly outperforms the ML. The only instance when the ML estimates appear to be closer to the population

	Quantiles				
λ	10	30	50	70	90
2	0.233	0.224	0.218	0.213	0.206
1/2	5.359	5.738	6.000	6.262	6.641
0	0.278	0.592	1.000	1.689	3.602
-1	-1.679	-1.392	-1.235	-1.102	-0.946

Table 1: Slopes of the population quantile functions for the Box-Cox regression model. The values for $\lambda = -1$ are multiplied by 100.

values is when $\lambda = 0$. For $\lambda = 2, 1/2, 1$, the QR median estimate is always closer to the mark, and the inter-quartile interval always contains the population slope.

This experiment is, admittedly, limited in scope. Nevertheless, it conveys a clear image of the flexibility of the linear QR estimator in adapting to a number of departures from non-linearity.

3 U.S. unemployment duration

3.1 Data

The unemployment duration data used in this exercise is the 1988 Displaced Worker Survey (DWS). The five-year, retrospective DWS has been conducted biannually since 1984 as a supplement to the Current Population Survey (CPS). Given its national representativeness and richness of information, the DWS supplements have been a major source of data for a burgeoning literature exploring the effects of displacement (e.g., Addison and Portugal, 1989; Farber, 1994; McCall, 1995).

The survey asks individuals from the regular CPS if, in any of the five years preceding the survey date, they had lost a job due to plant closing, an employer going out of business, a layoff from which the individual was not recalled, or other similar reasons. If the respondent has been displaced, he or she is asked a series of questions concerning the nature of the lost job and subsequent labor market experience, in particular, the time it took to get another job.

The DWS survey, unlike alternative administrative sources (e.g., Unemployment Insurance Registry), provides information on complete spells of joblessness. There are, nevertheless, incomplete spells (right-censoring) in this sample. On one hand, in a small number of cases, some individuals never found work following their dispacement. If they are engaged in search activity at the time of the survey, their spells of unemployment are obtained from the parent CPS. On the other hand, jobless spells are top-coded at 99 weeks. Overall, the proportion of censored observations is, in our sample, around 13 percent.

In this inquiry, because the nature of displacement is not well defined for certain individuals and sectors, those employed part-time and in agriculture at the time of displacement were excluded, as were those aged less than 20

		Quantiles				
λ		10	30	50	70	90
2	ML	0.239 0.233;0.245	0.229 0.224;0.235	0.223 0.218;0.228	0.218 0.214;0.223	0.211 0.207;0.215
	QR	0.233 0.225;0.241	0.224 0.218;0.231	0.219 0.213;0.224	0.213 0.208;0.219	0.206 0.199;0.213
1/2	ML	4.297 3.509;4.862	4.685 3.994;5.149	4.947 4.352;5.340	5.220 4.709;5.541	5.595 5.220;5.828
	QR	5.341 5.119;5.532	5.706 5.540;5.877	5.971 5.821;6.118	6.220 6.074;6.400	6.614 6.388;6.845
0	ML	0.278 0.268;0.286	0.592 0.572;0.610	1.000 0.967;1.031	1.689 1.634;1.742	3.602 3.483;3.714
	QR	0.268 0.247;0.288	0.553 0.523;0.582	0.904 0.858;0.953	1.476 1.399;1.558	2.932 2.763;3.124
-1	ML	-0.791 -0.814;-0.771	-0.900 -0.956;-0.866	-0.994 -1.085;-0.934	-1.103 -1.237;-1.018	-1.298 -1.522;-1.155
	QR	-0.662 -0.689;-0.634	-0.748 -0.772;-0.727	-0.820 -0.845;-0.797	-0.903 -0.928;-0.878	-1.045 -1.080;-1.009
1	ML	0.993 0.954;1.024	0.993 0.960;1.020	0.993 0-965;1.018	0.994 0.969;1.018	0.998 0.974;1.020
	QR	1.003 0.964;1.040	0.998 0.973;1.025	0.998 0.974;1.027	0.998 0.970;1.029	0.999 0.962;1.035

Table 2: ESTIMATED SLOPES OF THE QUANTILE FUNCTIONS FOR THE BOX-COX REGRESSION MODEL. The first entry reports the median value in 1000 replications for $\partial Q_Y(p|x)/\partial x$ for the indicated Box-Cox transformation parameter (λ) and p with the parameters estimated by maximum likelihood (ML) and by quantile regression (QR) of Y on X. The second entry is the inter-quartile range of these slope estimates. The values for $\lambda = -1$ are multiplied by 100.

and above 64 years of age of the end of the survey date. Similar reasoning explains the exclusion of the self-employed, together with those displaced for seasonal and "other reasons". Altogether, the restrictions imposed yielded an unweighted sample of 4076 individuals.

3.2 Estimation Procedures

Let $Q_p(y \mid x)$ for $p \in (0,1)$ denote the pth quantile of the distribution of the (log) unemployment duration, (y), given the vector, x, of covariates discussed above. The conditional quantile process – i.e., $Q_p(y \mid x)$ as a function of $p \in (0,1)$ – provides a full characterization of the conditional unemployment duration in much the same way as ordinary sample quantiles characterize a marginal distribution. We model these conditional quantiles as in (3) with link function $g(\cdot) = \exp(\cdot)$.

When there is no censoring, the quantile regression coefficients, $\beta(p)$, can be estimated for given $p \in (0,1)$ by the methods introduced by Koenker and Bassett (1978). Powell (1984, 1986) developed estimators of the QR coefficients for the case of censored data with known, but possibly varying, censoring points, (for a recent discussion of censored quantile regression see Fitzenberger, 1997). Consider a sample $(y_i, c_i, x_i),\ i = 1, \ldots, n$ where c_i denotes the upper threshold for y_i i.e., $y_i \leq c_i$ ($c_i = \infty$ when observation i is not censored). The QR estimator introduced by Powell minimizes the sample objective function

$$\sum_{i=1}^{n} \rho_p(y_i - \min[c_i, x_i'b])$$

with,

$$\rho_p(u) = \begin{cases} p\,u & \text{for } u \geq 0 \\ (p-1)\,u & \text{for } u < 0. \end{cases}$$

Estimation was performed iteratively using the LAD procedure in TSP. The iterative procedure is quite well known: at each iteration the observations with negative estimated residuals (i.e., those for which $x_i'\hat{\beta}(p) > c_i$) are discarded; then, the coefficients are re-estimated with the remaining observations until convergence is reached.

For the estimation of standard errors for the individual coefficients we resort to the bootstrap. This procedure was proposed by Bushinsky (1994) and named "ILPA". Since the "errors" from the QR equation are not necessarily homogeneously distributed, to achieve robustness we resample "triplets" (y, c, x) rather than the residuals from a particular QR fit. Notice also that, because of censoring, the threshold c must also be resampled.

3.3 Results

3.3.1 Quantile Regressions

Empirical results for selected quantiles from fitting the QR model are given in Table 1. For comparison purposes, we also provide the estimates obtained from a Cox proportional hazard model and from an accelerated failure time (AFT) model that employs an extended generalized gamma distribution.

In general, the regression coefficient estimates are fairly conventional. Age reduces escape rates, as does tenure in the previous job. Schooling enhances the chances of getting a job, whereas being unskilled decreases it. Higher state unemployment rates are associated with longer spells of joblessness. The familiar (opposing) effects of marital status on reemployment probabilities - positive for males and negative for females - are also obtained. It is also unsurprising that being non-white increases unemployment duration and being displaced by reason of the shutdown of the plant decreases duration. All these coefficients are statistically significant at conventional levels. Altogether less transparent are the effects of prenotification (both informal and written) and of previous wage rate on joblessness duration (but see below).

Comparison across different model specifications - Quantile Regression, Cox Proportional hazard, and Accelerated Failure Time - also reveals broad agreement, at least in terms of sign and statistical significance of the regression coefficients, in particular if we take the highest quantiles as comparators. The coefficient estimates for lower quantiles (for example, the 20th quantile in Table 1), however, disclose some interesting features. First, advance notice (both informal and written) of displacement exerts a significant influence on joblessness duration at low quantiles in contrast with the small and statistically insignificant effects at higher quantiles. Second, the impact of both the schooling and plant closing variables is much stronger at low quantiles. And third, the level of wages at the predisplacement job affects short spells of unemployment durations but not long ones. Clearly, these effects would not be detected by conventional parametric and semi-parametric approaches. Indeed, the results from the estimation of the AFT and Cox models appear to average out the time-varying regression effects.

Detailed quantile regression estimates are depicted in Figure 1. Coefficient estimates for $p \in (0.15, 0.85)$ are represented in conjunction with their corresponding confidence interval (The range of the confidence interval is plus or minus two bootstrap standard errors). To facilitate visual inspection, the median regression estimates are also represented by a horizontal line. The covariates seem to exert an influence that in most cases does not diverge from a constant effect over the entire conditional distribution. This agrees with a roughly constant goodness of fit indicator R^1(see Koenker and Machado, 1999). However, as hinted above, there are a number of exceptions, most notably, on the left tail—short-term unemployment — of the unemployment duration distribution.

The quantile equality tests provided in Table 2 give a more rigorous assessment of the constancy of regression coefficients for selected p (in this case, 0.2, 0.5, and 0.8). The indications provided by the tests are in line with the evolution of the regression coefficients exhibited in Figure 1. The tests suggest that the impact of some variables is short-term in nature: written and informal notice, education, and previous wage are in this category. Symmetrically, the impact of tenure in the previous job emerges late in spell of unemployment. The plant closing variable exhibits an effect that starts very strong in the early

	Quantile Regression				
	20th	50th	80th	Cox	AFT
Age (in years)	0.015 (0.006)	0.020 (0.004)	0.016 (0.003)	-0.013 (0.002)	0.021 (0.003)
Gender (male=1)	-0.004 (0.165)	0.111 (0.116)	0.082 (0.085)	-0.082 (0.057)	0.116 (0.087)
Race White=1	-.290 (0.158)	-.345 (0.099)	-0.370 (0.083)	0.322 (0.054)	-0.481 (0.082)
Marital status (married=1)	-0.323 (0.136)	-.214 (0.094)	-0.097 (0.072)	0.108 (0.046)	-0.189 (0.071)
Marital*Gender (married female=1)	0.495 (0.263)	0.625 (0.150)	0.451 (0.120)	-0.343 (0.074)	0.541 (.112)
Schooling (in years)	-0.080 (0.024)	-0.016 (0.016)	-0.019 (0.013)	0.018 (0.008)	-0.031 (0.012)
Tenure (in years)	-0.001 (0.013)	0.022 (0.008)	0.020 (0.005)	-0.009 (0.003)	0.014 (0.005)
Unskilled (Unskilled=1)	0.312 (0.122)	0.387 (0.086)	0.254 (0.065)	-0.200 (0.040)	0.330 (0.061)
Plant Closing (Shutdown=1)	-0.668 (0.123)	-0.357 (0.072)	-0.164 (0.057)	0.179 (0.034)	-0.321 (0.053)
Informal Notice (Notice=1)	-0.292 (0.123)	-0.081 (0.080)	-0.051 (0.057)	0.043 (0.035)	-0.082 (0.054)
Written Notice (Notice=1)	-0.757 (0.394)	0.097 (0.196)	0.031 (0.111)	-0.014 (0.078)	-0.038 (.120)
Unemp. Rate	0.0931 (0.026)	0.122 (0.018)	0.123 (0.014)	-0.076 (0.008)	0.116 (0.012)
Previous Wage (in logs)	-0.261 (0.119)	0.032 (0.077)	0.009 (0.069)	0.014 (0.037)	-0.038 (0.057)
Constant	1.167 (0.160)	2.432 (0.122)	3.598 (0.098)		2.890. (0.106)
Scale Parameter					1.565
Shape Parameter					0.613

Table 3: Unemployment Duration Regression Results (n=4076). The first entry in each cell is the regression coefficient point estimate; the second entry is estimated standard error; bootstrap standard errors from 1000 replications were obtained for the QR model; the parameterization of the extended generalized gamma distribution in the AFT model follows Addison and Portugal (1987); unemployment duration is in natural logs.

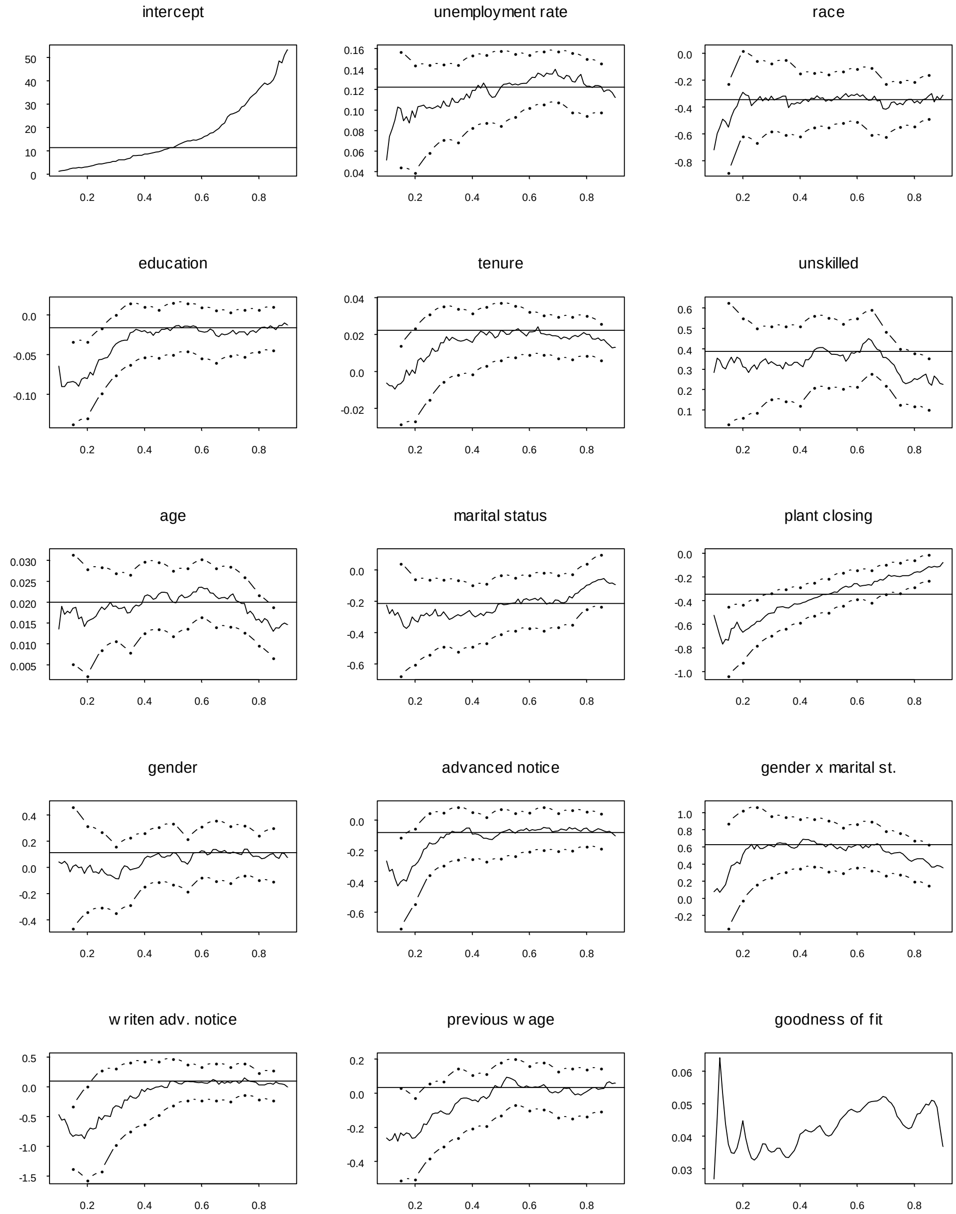

Figure 1: Quantile regression coefficients.The points represent 95% confidence intervals for $p = 0.15, 0.20, \ldots, 0.8$; the solid horizontal line represents the conditional median estimate. The intercept is plotted in natural units. The last panel plots the goodness of fit measure, (R^1), of Koenker and Machado (1999).

phase of the spell of unemployment and then fades away throughout. Some variables lose strength only in the late unemployment phase. This is the case of the Age and Unskilled covariates. In this model, solely three covariates appear to comply with the conventional homocedastic assumption: Race, Gender, and unemployment rate.

	Equality of Quantile Regression between:		
	20-50th Quant.	50-80th Quant.	20-80th Quant.
Age	-1.224	2.907	0.016
Gender	-0.712	-0.313	-0.673
Race	0.211	0.090	0.179
Marital status	-0.448	-2.818	-1.432
Marital*Gender	-.998	2.103	-0.243
Schooling	-2.903	0.554	-2.337
Tenure	-1.773	0.742	-1.409
Unskilled	0.726	3.869	0.534
Plant Closing	-2.065	-6.043	-3.569
Informal Notice	-2.026	-0.691	-2.044
Written Notice	-2.408	-0.339	-2.402
Unemp. Rate	-1.074	-0.424	-1.077
Previous Wage	-2.990	0.834	-2.285

Table 4: TESTS FOR THE STABILITY OF THE REGRESSION COEFFICIENTS AT SELECTED QUANTILES. Boostrap t-tests were obtained from 1000 draws for each quantile.

The tests provided in Table 2 are only suggestive, as they rely on specific choices for p. A preferable general approach is provided by the Wald test process suggested by Koenker and Machado (1999) which is graphically displayed in Figure 2. The information incorporated in the tests presented in Figure 2 enables a complete (and, thus, non-subjective) characterization of the effects of the covariates. Using the Wald process (and adequate critical values) it is possible to evaluate the significance of a given covariate on specific regions of the duration distribution. There are variables such as the unemployment rate, age and plant closing indicator, that exert a statistically significant influence throughout the entire distribution. On the other hand, gender and previous wage are not significant at any point. More interestingly, covariates such as education and pre-notification are only relevant on the left tail of the duration distribution, that is, for short-term unemployment.

It is worth noting that the variables that have significantly higher effects during the early phase of the unemployment spell very likely reflect the influence of on-the-job search (advance notice of displacement and dislocation by plant closing) or human capital (as captured by schooling and pre-displacement wage). In the latter case it can be argued that larger human capital endowments are associated with greater job opportunities and higher opportunity costs of unemployment that necessarily erode with the progression of the unemployment spell. A number of explanations can be suggested here. Human capital depreciation, unobserved individual heterogeneity correlated with the measures of human capital, or stigmatization would lead to a fading human capital effect on the transition rate out of unemployment.

It has been argued that the beneficial effects of prenotification accrue via the increase in on-the-job search intensity (Addison and Portugal, 1992). Faced with the prospect of an imminent discharge, the worker will engage in on-the-job search. If successful, he or she will experience a short spell of unemployment. Identically, workers displaced by reason of plant closing — in comparison with workers dismissed due to slack work or position shifted or abolished — benefit from an essentially short-term advantage conveyed by job search assistance and early (and unmistakable) warning of displacement.

In essence, both on-the-job search and human capital depreciation point to time varying effects of the covariates and, thus, to non-proportional hazard. These types of effects may be labeled "transient effects" after Cox and Oakes (1984).

3.3.2 Survival and Hazard Analysis

We have argued that the QR approach was flexible enough to enable the casting of its results in frames typical of alternative methodologies. Figure 3 presents estimates of the hazard and survivor functions for a reference "individual". Specifically, in the resampling procedure of Machado and Mata (2000) described in Section 2.2, the simulated sample was based on quantiles evaluated taking the sample means of the continuous covariates, and the reference category for the dummy explanatory variables.

The hazard function exhibits peaks at durations 4, 26, 39 and 52 weeks, due to the usual bunching of the answers as a result of the rounding of the measure of the jobless spells, mirroring beautifully the empirical hazard function. However, one should avoid being sidetracked by this heaping phenomenon in terms of the overall shape of the hazard function. A smoother graphic exhibition of the hazard function would show an inversed-U shaped function.

The variety of the time-varying effects on the covariates is depicted in Figure 4 with reference to the baseline hazard function. Figure 4 represents the effect of "unit" changes in the covariates on the log hazard function; for instance, the log hazard ratio for the (continuous) covariate j is

$$\log h(t|x_0 + \sigma_j e_j) - \log h(t|x_0)$$

where x_0 represents the reference vector of covariates described above, σ_j the sample standard deviation of covariate j and e_j is the jth unit vector. For the binary covariates the comparison is made between the two sub-populations. The results are directly comparable to those from the estimation of a Proportional hazard Model that are represented by the horizontal lines in Figure 4, (c.f., Table 3).

One conclusion is immediately apparent. Although the estimates from the PH model could be argued to be on average accurate, in most cases they pro-

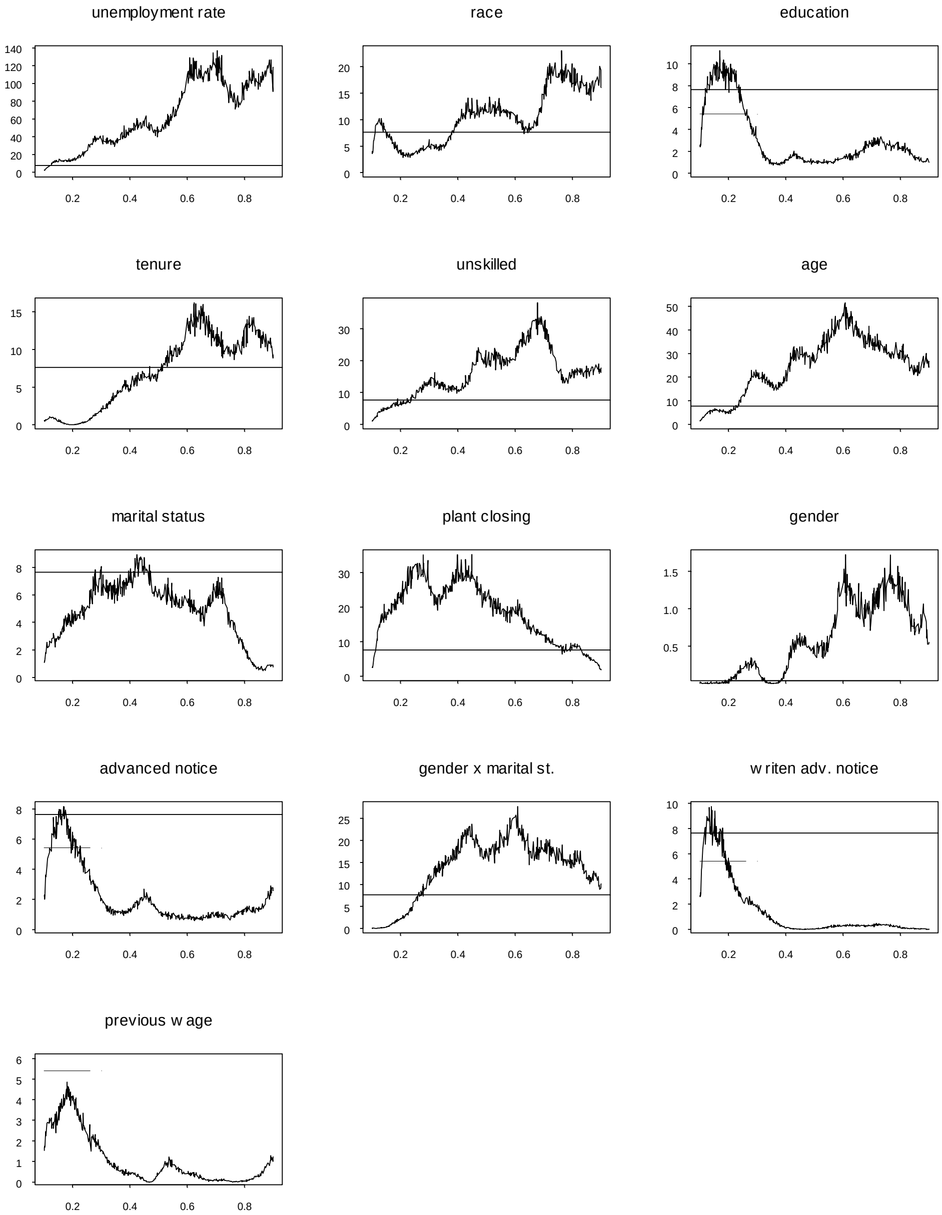

Figure 2: Wald test processes. Wald processes, $W(p)$, to test the significance of each QR coefficient as a function of $p \in (0.1, 0.9)$ (Koenker and Machado, 1999). The horizontal lines represent the 10% critical values of the $\sup_{p \in \Pi} W(p)$ statistic calculated by Andrews (1993); for the solid line $\Pi = (0.1, 0.9)$ and for the dashed line $\Pi = (0.1, 0.3)$.

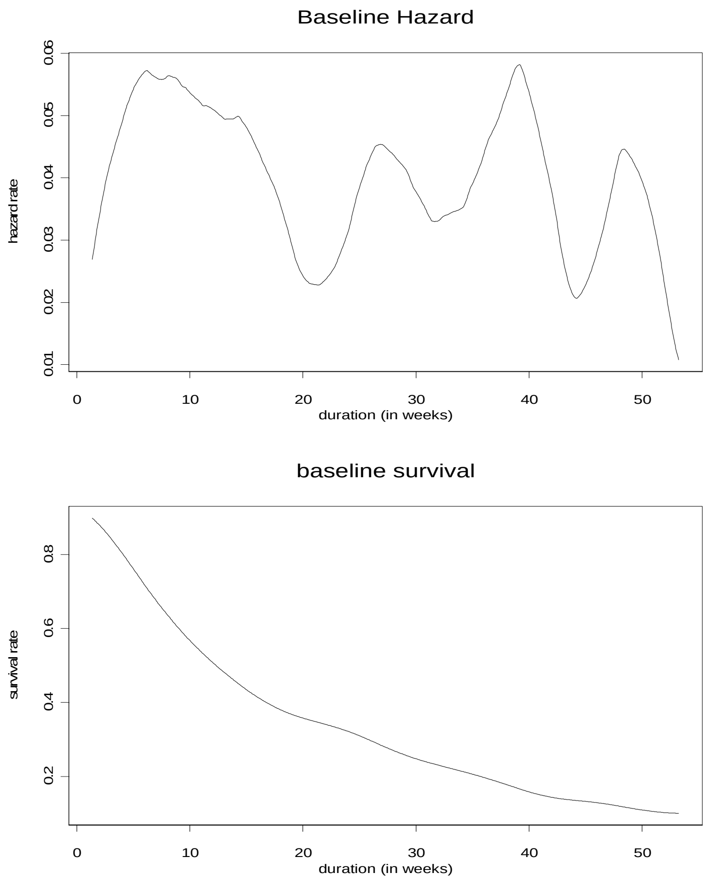

Figure 3: BASELINE HAZARD AND SURVIVOR FUNCTIONS. Hazard and Survivor functions implied by the QR evaluated at the sample means of the continuous covariates and at the reference sub-population for the discrete regressors.

vide only an oversimplified vision of the impact of the covariates on the exit rate from unemployment. For regressors such as "Advance notice", " Previous wage" or, even, "Education" and "Age", the PH estimates seem to provide a good approximation as the impact of those covaviates is roughly duration independent.

Some covariates, however, have impacts that are far from proportional. The impact of "Written Advance Notice" and "Plant Closing" are clearly decreasing with unemployment duration: the longer an individual stays unemployed, the smaller impact of these factors on the escape rate from unemployment. The detrimental impact of being male also increases with duration. On the other hand, unskilled workers' chances of leaving unemployment become less grim for those with longer spells.

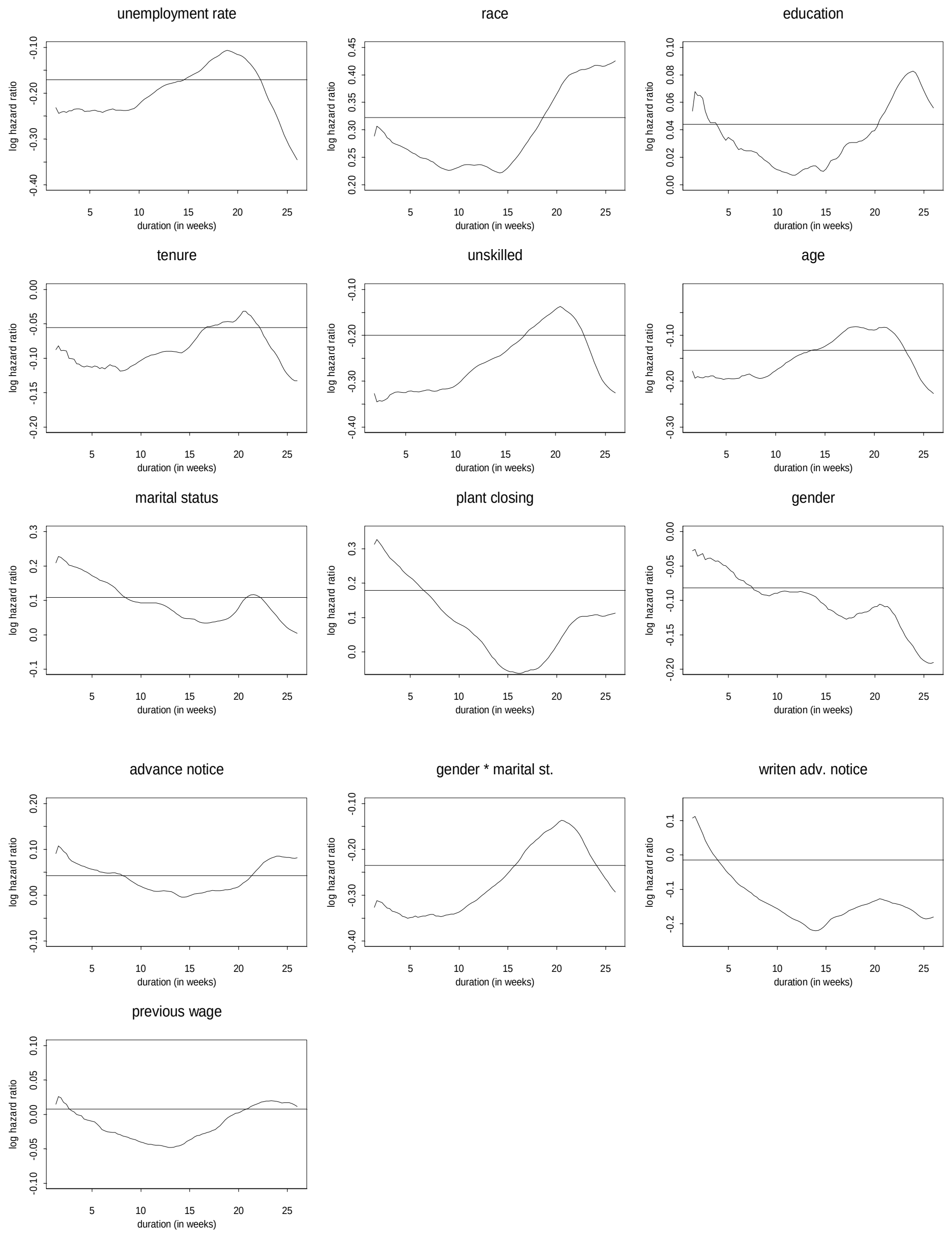

Figure 4: LOG HAZARD RATIOS. Log of the ratio of the hazard function for a sub-population with the corresponding characteristic or, for the continuous regressors, with the characteristic increased by one standard deviation, to the baseline hazard.

From a methodological vantage point, these results reveal that the hazard ratios estimated from models for conditional quantiles encompass the proportional hazard models as they allow sufficient flexibility for some regressors to have a proportional impact, while others depict effects that are duration dependent.

4 Conclusions

In this paper it is argued that quantile regression analysis offers a fruitful semi-parametric alternative for studying transition data. On one hand, the censored quantile regression estimator enables the accommodation of incomplete duration data. On the other hand, quantile regression naturally lends itself to the estimation of Accelerated Failure Time models, without imposing any distributional assumptions. Given the decreasing costs of computer-intensive statistical methods such as these, it is somewhat of a puzzle to realize that so few empirical studies have applied quantile regression models to duration data.

Apart from being a distribution-free model, there are other advantages accruing from the use of quantile regression models. First, it is a flexible approach in the sense that it allows the covariates to have different impacts at different points of the distribution. Second, the estimators of the regression coefficients are robust to the presence of (covariate uncorrelated) unobserved individual heterogeneity. Third, the estimators are resilient to misspecification of the functional form. And fourth, in comparison with conventional models, the quantile regression approach provides a much more complete characterization of the duration distribution.

It may be argued that a reason why researchers shy away from using the quantile regression estimator is its difficulty in dealing with standard survival analysis concepts. In this exercise it was shown, however, that it is straightforward to obtain typical survival outputs from quantile regression estimates (e.g., hazard and survival functions, mean residual life, conditional mean durations, etc.). Despite its robustness and flexibility, the use of quantile regression methods is not without some drawbacks. In particular, this approach is not well-suited to tackle sampling plans other than flow sampling, as in the case of the illustration presented here. In cases of stock sampling or sampling over a fixed time interval (see Lancaster, 1990) length-bias sampling issues are raised that can not be overcome straightforwardly within the framework of quantile regression methods. In these cases, maximum likelihood approaches may be prefered (Portugal and Addison, 2002).

Finally, in many instances, the quantile regression approach offers a natural and intuitive way to deal with some economic concepts. This is clearly the case of earnings inequality. It is, in our view, also the case of unemployment duration. In particular, the notions of short and long-term unemployment can be given an unambiguous empirical content. In the empirical illustration with U.S. unemployment duration, it was shown that some covariates impact differently at distinct regions of the unemployment duration distribution. The usefulness of the quantile regression approach is suggested by the conclusion that some variables impact solely at short durations (e.g., advance notice, schooling, and previous wage), impacts from other variables fade significantly over the course of the spell of unemployment (plant closing), while the effect of other variables remains constant across the board (gender and race). Those varying effects would remain undisclosed if conventional duration models were employed.

References

Addison, J. and P. Portugal (1987). On the distributional shape of unemployment duration. *Review of Economics and Statistics*, 68, 520-526.

Addison, J. and P. Portugal (1989). Job displacement, relative wage changes, and the duration of unemployment. *Journal of Labor Economics*, 7 281-302.

Addison, J. and P. Portugal (1992). Advance notice and unemployment: New evidence from the 1988 Displaced Worker Survey. *Industrial and Labor Relations Review*, 45, 645-664.

Andrews, D. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61, 821-856.

Bassett, G. and R. Koenker (1986). Strong consistency of regression quantiles and related empirical processes. *Econometric Theory*, 2, 191-201.

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: Application of quantile regression. *Econometrica*, 62, 405-458.

Buchinsky, M. (2001). Quantile regression with sample selection. Estimating women's return to education in the U. S. *Empirical Economics*, 26, 87-114.

Chamberlain, G. (1994). Quantile regression, censoring and the structure of wages. in C. Sims, eds. *Advances in Econometrics 6th World Congress vol. 1*, Cambridge Univ. Press.

Cox, D. and D. Oakes (1984). *Analysis of survival data.* Chapman and Hall.

Farber, H. (1999). Mobility and stability: the dynamics of job change in labor markets. in O. Ashenfelter Maddala and D. Card, eds. *Handbook of Labor Economics, vol. 3B*, North-Holland.

Fitzenberger, B. (1997). A guide to censored quantile regression. in G. Maddala and C. Rao, eds. *Handbook of Statistics, vol. 15*, Elsevier Science B. V.

Fitzenberger, B. (1998). The moving block bootstrap and robust inference for linear least squares and quantile regression. *Journal of Econometrics*, 82, 235-287.

Fitzenberger, B. and C. Kurz (1997). New insights on earning trends across skill groups and industries in West Germany. University of Konstanz.

Fitzenberger, B. Koenker, R. and Machado, J.A.F. (2001). Special issue on economic applications of quantile regression. *Empirical Economics*, 26.

Hendricks, W. and Koenker, R. (1992). Hierarquical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Society*, 87, 58-68.

Horowitz, J. and G. Neumann (1987). Semiparametric estimation of employment duration models. *Econometric Reviews*, 6,5-40.

Horowitz, J. and G. Neumann (1989). Specification testing in censored regression models: parametric and semiparametric methods. *Journal of Applied Econometrics*, 4, S61-S86.

Koenker, R. and G.S. Bassett Jr. (1978). Regression Quantiles. *Econometrica*, 46, 33-50.

Koenker, R. and J. Machado (1999). Goodness of fit and related inference processes for quantile regresssion. *Journal of the American Statistical Association*, 94, 1296-1310.

Koenker, R. and O. Geling (2001). Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association*, 96, 458-468.

Koenker, R. and Y. Billias (2001). Quantile regression for duration data: a reappraisal of the Pennsylvania reemployment bonus experiments. *Empirical Economics*, 26, 199-220.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data.*Cambridge University Press.

Machado, J. and J. Mata (2000). Counterfactual decomposition of changes in wage distributions using quantile regression. Preprint.

Machado, J. and J. Mata (2001). Earning functions in Portugal 1982-1994: evidence from quantile regressions. *Empirical Economics*, 26, 115-134.

Manski, C. (1988). *Analog estimation methods in econometrics.* Chapman and Hall.

McCall, B. (1996). Unemployment insurance rules, joblessness, and part-time work.*Econometrica*, 64, 647-682.

Portugal,P. and J. Addison (1995). Short and long-term unemployment: a parametric model with time-varying effects. *Oxford Bulletin of Economics and Statistics*, 57, 207-227.

Portugal,P. and J. Addison (2002). Unemployment duration: Competing and defective risks. *Journal of Human Resources*, forthcoming.

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

Powell, J. (1986). Censored regression quantiles. *Journal of Econometrics*, 32, 143-155.

Ridder, G., (1987). The sensitivity od duration models to misspecified unobserved heterogeneity and duration dependence. Manuscript.

Silverman, (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.