Fitzenberger, Bernd; Wilke, Ralf A.

Working Paper

Using Quantile Regression for Duration Analysis

ZEW Discussion Papers, No. 05-65

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Fitzenberger, Bernd; Wilke, Ralf A. (2005) : Using Quantile Regression for Duration Analysis, ZEW Discussion Papers, No. 05-65, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim

This Version is available at:
https://hdl.handle.net/10419/24161

Discussion Paper No. 05-65

Using Quantile Regression for Duration Analysis

Bernd Fitzenberger and Ralf A. Wilke

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Centre for European Economic Research

Discussion Paper No. 05-65

Using Quantile Regression for Duration Analysis

Bernd Fitzenberger and Ralf A. Wilke

Non–technical Summary

This survey summarizes recent estimation approaches using quantile regression for (right-censored) duration data. We provide a discussion of the advantages and drawbacks of quantile regression in comparison to popular alternative methods such as the (mixed-)proportional hazard model or the accelerated failure time model. We argue that quantile regression methods are robust and flexible in a sense that they can capture a variety of effects at different quantiles of the duration distribution. This survey discusses a selection of well known theoretical results and adds some new theoretical insights on the relationship between the proportional hazard rate model, the presence of unobserved heterogeneity, and the estimation of quantile regression. Our discussion emphasizes cases when evidence based on quantile regression implies a rejection of the proportional hazard model. Quantile regression also allow to estimate hazard rates which are often of interest in duration analysis and the method can be extended to a nonlinear Box-Cox transformation of the duration variable. Furthermore, the survey points out that quantile regression can not take account of time–varying covariates and it has not been extended so far to account for unobserved heterogeneity and competing risks.

We illustrate our theoretical reasoning by an application of the quantile regression to unemployment duration data for young workers in West-Germany. We find that some variables change the direction of their influence on unemployment duration across the distribution. This implies that the basic assumptions for a proportional hazard model is violated in our application. These empirical findings motivate the use of the estimation approaches discussed in this survey. Our results indicate that the overall unemployment rate shows no impact on the distribution of unemployment durations and that the duration of unemployment for young workers has become shorter over time.

Using Quantile Regression for Duration Analysis*

Bernd Fitzenberger† and Ralf A. Wilke‡

August 2005

*This is a longer version of our paper published (in: *Allgemeines Statistisches Archiv* 90(1)). Here, we add the discussion of Box-Cox quantile regression and the details of the empirical application. The paper benefitted from the helpful comments by an anonymous referee. We gratefully acknowledge financial support by the German Research Foundation (DFG) through the research project "Microeconometric modelling of unemployment durations under consideration of the macroeconomic situation". Thanks are due to Eva Müller and Xuan Zhang for excellent research assistance. The empirical work in this paper uses partly the IAB employment subsample 1975-2001 - regional file - which is a 2% random sample of all employees in Germany who have been covered by the social insurance system for at least one day in the period under observation. This period lasts from 1975 to 2001 in West Germany and from 1992 to 2001 in East Germany. The file includes as well information on the receipt of any kind of unemployment compensation from the German Federal Employment Service (FES) during this period. The sample is extracted from the "Beschaeftigten-Leistungsempfnger-Historik (BLH)" (Employees and benefits recipients history file) of the Institute for Employment Research (IAB) which is a part of the FES. The IAB does not take any responsibility for the use of its data. All errors are our sole responsibility.

†Corresponding author: Goethe University Frankfurt, ZEW, IZA, IFS. Address: Department of Economics, Goethe-University, PO Box 11 19 32 (PF 247), 60054 Frankfurt am Main, Germany. E-mail: fitzenberger@wiwi.uni-frankfurt.de

‡Zentrum für Europäische Wirtschaftsforschung (ZEW), P.O. Box 10 34 43, 68034 Mannheim, Germany. E–mail: wilke@zew.de.

Abstract

Quantile regression methods are emerging as a popular technique in econometrics and biometrics for exploring the distribution of duration data. This paper discusses quantile regression for duration analysis allowing for a flexible specification of the functional relationship and of the error distribution. Censored quantile regression address the issue of right censoring of the response variable which is common in duration analysis. We compare quantile regression to standard duration models. Quantile regression do not impose a proportional effect of the covariates on the hazard over the duration time. However, the method can not take account of time–varying covariates and it has not been extended so far to allow for unobserved heterogeneity and competing risks. We also discuss how hazard rates can be estimated using quantile regression methods. A small application with German register data on unemployment duration for younger workers demonstrates the applicability and the usefulness of quantile regression for empirical duration analysis.

Keywords: censored quantile regression, unemployment duration, unobserved heterogeneity, hazard rate

JEL: C13, C14, J64

Contents

1 Introduction **1**

2 Quantile Regression and Duration Analysis **2**

2.1 Quantile Regression and Proportional Hazard Rate Model 2

2.2 Censoring and Censored Quantile Regression 5

2.3 Estimating the Hazard Rate based on Quantile Regression 7

2.4 Box–Cox Quantile Regression 9

2.5 Unobserved Heterogeneity 11

3 Unemployment Duration among Young Germans **14**

3.1 Data and Institutions 15

3.2 Estimation Results 18

4 Summary **22**

Appendix **27**

1 Introduction

Duration data are commonly used in applied econometrics and biometrics. There is a variety of readily available estimators for popular models such as the accelerated failure time model and the proportional hazard model, see e.g. Kiefer (1988) and van den Berg (2001) for surveys. Quantile regression is recently emerging as an attractive alternative to these popular models (Koenker and Bilias, 2001; Koenker and Geling, 2001; Portnoy; 2003). By modelling the distribution of the duration in a flexible semiparametric way, quantile regression do not impose modelling assumptions that may not be empirically valid, e.g. the proportional hazard assumption. Quantile regression are more flexible than accelerated failure time models or the Cox proportional hazard model because they do not restrict the variation of estimated coefficients over the quantiles. Estimating censored quantile regression allows to take account of right censoring which is present in typical applications of duration analysis (Powell, 1984; Fitzenberger, 1997). However, quantile regression involve three major disadvantages. First, the method is by definition restricted to the case of time–invariant covariates. Second, there is no competing risks framework yet and third, so far quantile regression does not account for unobserved heterogeneity, which is a major ingredient of the mixed proportional hazard rate model.

Quantile regression model the changes of quantiles of the conditional distribution of the duration in response to changes of the covariates. In actual applications of duration analysis, researchers are often interested in the effects on the hazard rate after a certain elapsed duration and how the hazard rate changes with the elapsed duration (duration dependence). Machado and Portugal (2002) and Guimarães et al. (2004) have introduced a simple simulation method to obtain the conditional hazard rates implied by the quantile regression estimates. In this paper, we present a slightly modified version of their estimator. The modifications are necessary to overcome difficulties in the case of censored data and to fix a general smoothing problem. Using this method, it is straightforward to analyze duration dependence without having to assume that the pattern estimated for the so-called baseline hazard in proportional hazard rate models applies uniformly to all observations with different covariates.

Section 2 discusses important aspects of quantile regression methods for duration analysis and shows how conditional hazard rates can be obtained from estimated quantile regression coefficients. Section 3 presents an application of censored quantileregression to German unemployment duration data that demonstrates the usefulness of the applied methods for empirical research. Section 4 summarizes.

2 Quantile Regression and Duration Analysis

This section discusses quantile regression as an econometric tool to estimate duration models and addresses various issues involved. Quantile regression are contrasted with the popular proportional hazard rate model. The section provides a short survey on the important methodological aspects, not all of them will be addressed in the subsequent empirical application.

2.1 Quantile Regression and Proportional Hazard Rate Model

Koenker und Bassett (1978) introduced quantile regressions[1] as a regression based method to model the quantiles of the response variable conditional on the covariates. Our focus is on linear quantile regression for duration data involving the estimation of the accelerated failure time model at different quantiles $\theta \in (0,1)$ for the completed duration T_i of spell i

$$h(T_i) = x_i'\beta^\theta + \epsilon_i^\theta, \tag{1}$$

where the θ-quantile of ϵ_i^θ conditional on x_i, $q_\theta(\epsilon_i^\theta|x_i)$, is zero and $h(.)$ is a strictly monotone transformation preserving the ordering of the quantiles. The most popular choice is the log–transformation $h(.) = log(.)$. The transformation can either be chosen a priori (e.g. as being the log–transformation) or it can be estimated by choosing among a class of transformation functions (e.g. among the set of possible Box–Cox–transformations, see e.g. Buchinsky, 1995, or Machado and Mata, 2000, as discussed in section 2.4). Due to the invariance of quantiles under positive monotone transformations, quantile regression are not restricted to a linear specification of the conditional quantiles. In fact, quantile regression model the conditional quantile of the response variable $q_\theta(h(T_i)|x_i) = x_i'\beta^\theta$ or, alternatively, due to the invariance property of quantiles $q_\theta(T_i|x_i) = h^{-1}(x_i'\beta^\theta)$. Modelling conditional quantiles is an indirect way to model the conditional distribution function of $log(T_i)$ given x_i. The linear specification allows for differences in the impact of covariates x_i across the conditional distribution of the response variable. However, the specification imposes that the coefficient is the same for a given quantile θ irrespective of the actual value of $q_\theta(h(T_i)|x_i)$.

We will discuss the asymptotic distribution for linear quantile regression in the next subsection for the case of censored quantile regression which nests the case without censoring. The asymptotic distribution in the case with a smooth transformation function $h(.)$ depending on unknown parameters to be estimated can be found in Powell (1991), Chamberlain (1994), or Fitzenberger et al.

[1] See Buchinsky (1998) and Koenker and Hallock (2002) for surveys. The collection of papers in Fitzenberger, Koenker and Machado (2001) comprises a number of economic applications of quantile regressions, among others, the paper by Koenker and Bilias (2001) using quantile regression for duration analysis.

(2004), who treat the special case of Box–Cox transformation. The asymptotic results generalize in a straight forward manner to other smooth transformation functions.

A possible problem of quantile regression is the possibility that the coefficient estimates can be quite noisy (even more so for censored quantile regression) and often non-monotonic across quantiles. To mitigate this problem, it is possible to obtain smoothed estimates through a minimum–distance approach. One can investigate, whether a parsimonious relation describes the movement of the coefficients across quantiles by minimizing the following quadratic form $(\hat{\beta} - f[\delta])'\hat{\Psi}^{-1}(\hat{\beta} - f[\delta])$ with respect to δ, the coefficients of a smooth parametric specification of the coefficients as a funtion of θ. $\hat{\beta}$ is the stacked vector of quantile regression coefficient estimates $\hat{\beta}^{\theta}$ at different quantiles and $\hat{\Psi}$ is the estimated covariance matrix of $\hat{\beta}$, see next subsection for the asymptotic distribution. This approach is not pursued in the application below. We are not aware of any application of this approach in the literature.

The most popular parametric Cox proportional hazard model (PHM), Kiefer (1988), is based on the concept of the hazard rate conditional upon the covariate vector x_i given by

$$\lambda_i(t) = \frac{f_i(t)}{P(T_i \geq t)} = exp(x_i'\tilde{\beta})\lambda_0(t) \ , \tag{2}$$

where $f_i(t)$ is the density of T_i at duration t and $\lambda_0(t)$ is the so called baseline hazard rate. The hazard rate is the continuous time version of an instantaneous transition rate, i.e. it represents approximately the conditional probability that the spell i ends during the next period of time after t conditional upon survival up to period t.

There is a one–to–one correspondence between the hazard rate and the survival function, $S_i(t) = P(T_i \geq t)$, of the duration random variable, $S_i(t) = exp\left(-\int_0^t \lambda_i(s)ds\right)$. A prominent example of the parametric[2] proportional hazard model is the Weibull model where $\lambda_0(t) = pt^{p-1}$ with a shape parameter $p > 0$ and the normalizing constant is included in $\tilde{\beta}$. The case $p = 1$ is the special case of an exponential model with a constant hazard rate differentiating the increasing $(p > 1)$ and the decreasing $(0 < p < 1)$ case. Within the Weibull family, the proportional hazard specification can be reformulated as the accelerated failure time model

$$log(T_i) = x_i'\beta + \epsilon_i \tag{3}$$

where $\beta = -p^{-1}\tilde{\beta}$ and the error term ϵ_i follows an extreme value distribution, Kiefer (1988, sections IV and V).

The main thrust of the above result generalizes to any PHM (2). Define the integrated baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(\tilde{t})d\tilde{t}$, then the following well known generalization of the accelerated failure

[2]Cf. Kiefer(1988, section III.A) for nonparametric estimation of the baseline hazard $\lambda_0(t)$.

time model holds

$$log(\Lambda_0(T_i)) = x_i'\beta + \epsilon_i \tag{4}$$

with ϵ_i again following an extreme value distribution and $\beta = -\tilde{\beta}$, see Koenker and Bilias (2001) for a discussion contrasting this result to quantile regression. Thus, the proportional hazard rate model (2) implies a linear regression model for the a priori unknown transformation $h(T_i) = log(\Lambda_0(T_i))$. This regression model involves an error term with an a priori known distribution of the error term and a constant coefficient vector across quantiles.

From a quantile regression perspective, it is clear that these properties of the PHM are quite restrictive. Provided the correct transformation is applied, it is possible to investigate whether these restrictions hold by testing for the constancy of the estimated coefficients across quantiles. Testing whether the error term follows an extreme value distribution is conceivable though one has to take account of possible shifts and normalizations implied by the transformation. However, if a researcher does not apply the correct transformation in (4), e.g. the log transformation in (3) is used though the baseline hazard is not Weibull, then the implications are weaker. Koenker and Geling (2001, p. 462) show that the quantile regression coefficients must have the same sign if the the data is generated by a PHM.

A strong, and quite apparent violation of the proportional hazard assumption occurs, if for two different covariate vectors x_i and x_j, the survival functions $S_i(t)$ and $S_j(t)$, or equivalently the predicted conditional quantiles, do cross. Crossing occurs, if for two quantiles $\theta_1 < \theta_2$ and two values of the covariate vector x_i and x_j, the ranking of the conditional quantiles changes, e.g. if $q_{\theta_1}(T_i|x_i) < q_{\theta_1}(T_j|x_j)$ and $q_{\theta_2}(T_i|x_i) > q_{\theta_2}(T_j|x_j)$. Our empirical application below involves cases with such a finding. This is a valid rejection of the PHM, provided our estimated quantile regression provides a sufficient goodness–of–fit for the conditional quantiles.

There are three major advantages of PHMs compared to quantile regressions as discussed in the literature. PHMs can account for unobserved heterogeneity, for time varying covariates, and for competing risks in a straight forward way (Wooldridge, 2002, chapter 20). The issue of unobserved heterogeneity will be discussed at some length below. The estimation of competing risks models with quantile regression has not been addressed in the literature. This involves a possible sample selection bias, an issue which has only be analyzed under much simpler circumstances for quantile regression (Buchinsky, 2001). In fact, this is a dynamic selection problem which, also in the case of a PHM, requires fairly strong identifying assumptions.

It is natural to consider time varying covariates when the focus of the analysis is the hazard rate as a proxy for the exit rate during a short time period. This is specified in a PHM as

$$\lambda_i(t) = exp(x_{i,t}'\tilde{\beta})\lambda_0(t) \tag{5}$$

and there are readily available estimators for this case. It is not possible anymore to transform this model directly into an accelerated failure time model which could be estimated by regression methods.

Assuming strict exogeneity of the covariates, it is straight forward to estimate proportional hazard models with time varying coefficients (Wooldridge, 2002, chapter 20). If under strict exogeneity the complete time path of the covariates is known, it is conceivable – though often not practical – to condition the quantile regression on the entire time path to mimick the time varying effect of the covariates. A natural example in the analysis of unemployment durations would be that eligibility for unemployment benefits is exhausted after a certain time period and this is known ex ante. In fact, in such a case quantile regression also naturally allow for anticipation effects which violates specification (5). In many cases, the time path of time–varying covariates is only defined during the duration of the spell, which is referred to as internal covariates (Wooldridge, 2002, p. 693). Internal covariates typically violate the strict exogeneity assumption and it is difficult to relax the strict exogeneity assumption when also accounting for unobserved heterogeneity.

The case of time varying coefficients β_t can be interpreted as a special case of time–varying covariates by interacting the covariates with dummy variables for different time periods. However, if the specification of the baseline hazard function is very flexible then an identification issue can arise. Time varying coefficients β_t are similar in spirit to quantile regressions with changing coefficients across conditional quantiles. While the former involves coefficients changes according to the actual elapsed duration, the latter specifies these changes as a function of the quantile. It depends on the application as to which approach can be better justified.

Summing up the comparison so far, while there are some problems when using the PHM with both unobserved heterogeneity and time–varying covariates, the PHM can take account of these issues in a somewhat better way than quantile regression. Presently, there is also a clear advantage of the PHM regarding the estimation of competing risk models. However, the estimation of a PHM comes at the cost of the proportional hazard assumption which itself might not be justifiable in the context of the application of interest.

2.2 Censoring and Censored Quantile Regression

Linear censored quantile regression, introduced by Powell (1984, 1986), allow for semiparametric estimation of quantile regression for a censored regression model in a robust way. A survey on the method can be found in Fitzenberger (1997). Since only fairly weak assumptions on the error terms are required, censored quantile regression (CQR) is robust against misspecification of the error term. Horowitz and Neumann (1987) were the first to use CQR's as a semiparametric

method for an accelerated failure time model of employment duration.

Duration data are often censored. Right censoring occurs when we only observe that a spell has survived until a certain duration (e.g. when the period of observation ends) but we do not know exactly when it ends. Left censoring occurs when spells observed in the data did start before the beginning of the period of observation. Spells who started at the same time and who finished before the beginning of the period of observation are not observed. Quantile regression can not be used with left censored data.[3] Left censoring is also difficult to handle for PHMs since strong assumptions have to be invoked to estimate the model. In the following, we only consider the case of right censoring which both PHM and CQR are well suited for. Thus, we can only analyze so–called flow samples (Wooldridge, 2002, chapter 20) of spells for which the start of the spells lies in the time period of observation.[4]

Let the observed duration be possibly right censored in the flow sample, i.e. the observed completed duration T_i is given by $T_i = min\{T_i^*, yc_i\}$, where T_i^* is the true duration of the spell and yc_i is the spell specific threshold value (censoring point) beyond which the spell can not be observed. For the PHM, this can be incorporated in maximum likelihood estimation analogous to a censored regression model (Wooldridge, 2002, chapter 20) and it is not necessary to know the potential censoring points yc_i for uncensored observations. In contrast, CQR requires the knowledge of yc_i irrespective of whether the observation is right censored. CQR provide consistent estimates of the quantile regression coefficients β^θ in the presence of fairly general forms of fixed censoring.[5] The known censoring points can either be deterministic or stochastic and they should not bunch in a certain way on or around the true quantile regression line, see the discussion in Powell (1984).

Estimating linear CQR involves minimizing the following distance function

$$\sum_{i=1} \rho_\theta(ln(T_i) - min(x_i'\beta^\theta, yc_i)) \tag{6}$$

with respect to β^θ, where the so–called "check function" $\rho_\theta(z) = \theta \cdot |z|$ for $z \geq 0$ and $\rho_\theta(z) = (1-\theta) \cdot |z|$ for $z < 0$ and yc_i denotes the known observation specific censoring points. A quantile regression without censoring is nested as the special case with $yc_i = +\infty$.

[3]Two–limit censored quantile regression (Fitzenberger, 1997) can be used in the rare situation when all spells are observed which start before the start of the observation period and, in case they end before the start of the observation period, the exact length of the spell is not known.

[4]Analyzing all spells observed at some point of time during the period of observations involves a so–called stock sample also including left–censored observations.

[5]Refer to Buchinsky and Hahn (1998) for a semiparametric extension of CQR to the case when the censoring points are not known for the uncensored observations (random censoring).

Powell (1984, 1986) showed that the CQR estimator $\hat{\beta}^{\theta}$ is $\sqrt{N}$–consistent and asymptotically normally distributed, see also Fitzenberger (1997) for a detailed discussion of the asymptotic distribution. A crucial feature of this result is that the asymptotic distribution depends only upon those observations where the fitted quantiles are not censored, i.e. $I(x_i'\hat{\beta}^{\theta}<yc_i)=1$.

The actual calculation of the CQR–estimator based on individual data is numerically very difficult, since the distance function (6) to be minimized is not convex. This is in contrast to quantile regression without censoring. There are a number of procedures suggested in the literature to calculate the CQR–estimator (Buchinsky, 1998, Fitzenberger, 1997, and Fitzenberger and Winker, 2001).[6]

For heteroscedasticity–consistent inference, researchers often resort to bootstrapping, see e.g. Buchinsky (1998) and Fitzenberger (1997, 1998), using the Design–Matrix–Bootstrap (often also denoted as "pairwise bootstrap"). The covariance of the CQR estimates across quantiles can easily be estimated by basing those estimates on the same resample. Bilias et al. (2000) suggest a simplified version of the bootstrap for CQR by showing that it suffices asymptotically to estimate a quantile regression without censoring in the resample based only on those observations for which the fitted quantile is not censored, i.e. $x_i'\hat{\beta}^{\theta} < yc_i$.

In the empirical application below we show that several estimated coefficients change their sign across quantiles and therefore they do not support empirically the proportional hazard model.

2.3 Estimating the Hazard Rate based on Quantile Regression

Applications of duration analysis often focus on the impact of covariates on the hazard rate. Quantile regression estimate the conditional distribution of T_i conditional on covariates and it is possible to infer the estimated conditional hazard rates from the quantile regression estimates.

A direct approach is to construct a density estimate from the fitted conditional quantiles $\hat{q}_{\theta}(T_i|x_i) = h^{-1}(x_i'\hat{\beta}^{\theta})$. A simple estimate for the hazard rate as a linear approximation of the hazard rates between the different θ–quantiles would be

$$\hat{\lambda}_i(t) = \frac{(\theta_2 - \theta_1)}{\left(h^{-1}(x_i'\hat{\beta}^{\theta_2}) - h^{-1}(x_i'\hat{\beta}^{\theta_1})\right)(1 - 0.5(\theta_1 + \theta_2))}, \tag{7}$$

where $\hat{\lambda}_i(t)$ approximates the hazard rates between the estimated θ_1–quantile and θ_2–quantile.[7] Two points are noteworthy. First, the estimated conditional quantiles are piecewise constant due

[6] In light of the numerical difficulties, a number of papers have, in fact, suggested to change the estimation problem to make it convex (e.g. Buchinsky and Hahn, 1998, Chernozhukov and Hong, 2002, and Portnoy, 2003).

[7] A similar estimator based on the estimate of the sparsity function is described in Machado and Portugal (2002). It shares the same problems discussed for the estimator presented in (7).

to the linear programming nature of quantile regression (Koenker and Bassett, 1978, Fitzenberger, 1997). Second, it is not guaranteed that the estimated conditional quantiles are ordered correctly, i.e. it can occur that $\hat{q}_{\theta_1}(T_i|x_i) > \hat{q}_{\theta_2}(T_i|x_i)$ even though $\theta_1 < \theta_2$. Therefore, θ_1 and θ_2 have to be chosen sufficiently far apart to guarantee an increase in the conditional quantiles.

In order to avoid these problems, Machado and Portugal (2002) and Guimarães et al. (2004) suggest a resampling procedure (henceforth denoted as GMP) to obtain the hazard rates implied by the estimated quantile regression. The main idea of GMP is to simulate data based on the estimated quantile regressions for the conditional distribution of T_i given the covariates and to estimate the density and the distribution function directly from the simulated data.

In detail, GMP works as follows (see Machado and Portugal, 2002, and Guimarães et al., 2004), possibly only simulating non–extreme quantiles:

1. Generate M independent random draws $\theta_m, m = 1, ..., M$, from a uniform distribution on (θ_l, θ_u), i.e. extreme quantiles with $\theta < \theta_l$ or $\theta > \theta_u$ are not considered here. θ_l and θ_u are chosen in light of the type and the degree of censoring in the data. Additional concerns relate to the fact that quantile regression estimates at extreme quantiles are typically statistically less reliable, and that duration data might exhibit a mass point at zero or other extreme values. The benchmark case with the entire distribution is given by $\theta_l = 0$ and $\theta_u = 1$.

2. For each θ_m, estimate the quantile regression model obtaining M vectors β^{θ_m} (and the transformation $h(.)$ if part of the estimation approach).

3. For a given value of the covariates x_0, the sample of size M with the simulated durations is obtained as, $T_m^* \equiv \hat{q}_{\theta_m}(T_i|x_0) = h^{-1}(x_0'\hat{\beta}^{\theta_m})$ with $m = 1, ..., M$.

4. Based on the sample $\{T_m^*, m = 1, ..., M\}$, estimate the conditional density $f^*(t|x_0)$ and the conditional distribution function $F^*(t|x_0)$.

5. We suggest to estimate the hazard rate conditional on x_0 in the interval (θ_l, θ_u) by[8]

$$\hat{\lambda}_0(t) = \frac{(\theta_u - \theta_l) f^*(t|x_0)}{1 - \theta_l - (\theta_u - \theta_l) F^*(t|x_0)} \quad .$$

   Simulating the full distribution ($\theta_l = 0$ and $\theta_u = 1$), one obtains the usual expression: $\hat{\lambda}_0(t) = f^*(t|x_0)/[1 - F^*(t|x_0)]$.

[8] $f^*(t|x_0)$ estimates the conditional density in the quantile range (θ_l, θ_u), i.e. $f(t|q_{\theta_l}(T|x_0) < T < q_{\theta_u}(T|x_0), x_0)$, and the probability of the conditioning event is $\theta_u - \theta_l = P(q_{\theta_l}(T|x_0) < T < q_{\theta_u}(T|x_0)|x_0)$. By analogous reasoning, the expression in the denominator corresponds to the unconditional survival function, see Zhang (2004) for further details.

This procedure (step 3) is based on the probability integral transformation theorem from elementary statistics implying $T_m^* = F^{-1}(\theta_m)$ is distributed according to the conditional distribution of T_i given x_0 if $F(.)$ is the conditional distribution function and θ_m is uniformly distributed on $(0, 1)$. Furthermore, the fact is used that the fitted quantile from the quantile regression provides a consistent estimate of the population quantile, provided the quantile regression is correctly specified.

The GMP procedure uses a kernel estimator for the conditional density

$$f^*(t|x_0) = 1/(M\,h) \sum_{m=1}^{M} K\left((t - T_i^*)/h\right)$$

where h is the bandwidth and $K(.)$ the kernel function. Based on this density estimate, the distribution function estimator is $F^*(t|x_0) = 1/M \sum_{m=1}^{M} \mathcal{K}\left((t - T_i^*)/h\right)$ with $\mathcal{K}(u) = \int_a^t K(v)\,dv$. Machado and Portugal (2002) and Guimarães et al. (2004) suggest to start the integration at zero ($a = 0$), probably because durations are strictly positive. However, the kernel density estimator also puts probability mass into the region of negative durations, which can be sizeable with a large bandwidth, see Silverman (1986, section 2.10). Using the above procedure directly, it seems more advisable to integrate starting from minus infinity, $a = -\infty$. A better and simple alternative would be to use a kernel density estimator based on log durations. This is possible when observed durations are always positive, i.e. there is no mass point at zero. Then, the estimates for density and distribution function for the duration itself can easily be derived from the density estimates for log duration by applying an appropriate transformation.[9]

In our empirical application below we find that this flexible econometric method can reveal interesting results that would not show up under stronger conditions. Since the estimated hazard rates even cross, a proportional hazard model is inappropriate in this case.

2.4 Box–Cox Quantile Regression

The linear specification of the quantile regression is less restrictive as it may seem since because of the rank preservation of the response variable under strictly positively monotone transformation $h(.)$, as in equation (1), implying $q_\theta(h(T_i)|x_i) = h(q_\theta(T_i|x_i))$. A fairly flexible extension of linear quantile regression as a single index model is given by jointly estimating the transformation $h(.)$ together with the regression coefficients β. Such a model is also suitable for the generalization of the accelerated failure time model in equation (4). Horowitz (1996) and Chen (2002) suggest nonparametric estimators for the transformation $h(.)$ given estimates of the unknown coefficients

[9] Silverman (1986, section 2.10) discusses further alternatives for this problem.

β. Their estimators poses formidable problems in finite samples, in particular, when the joint density of (T, x) is low. Not being practical at this point, their estimation strategy is not pursued any further here.

A popular parametric choice for the transformation of the dependent variable is introduced by Box and Cox (1964):

$$h(T) = T_{\lambda,i} = \begin{cases} (T_i^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ log(T_i) & \text{if } \lambda = 0, \end{cases}$$

where λ can take any value in $\mathcal{R}$. Thus, we obtain a linear model for $q_\theta(T_{\lambda,i}|x) = x_i'\beta^\theta$ and the quantile for T_i can be written as

$$q_\theta(T_i|x) = (\lambda x_i'\beta^\theta + 1)^{1/\lambda} \quad . \tag{8}$$

The possibility to estimate λ allows for flexibility in estimating the model in (1). Powell (1991), Chamberlain (1994), Buchinsky (1995), Machado and Mata (2000), and recently Fitzenberger et al. (2004) provide further details on Box–Cox quantile regression. The estimator is $\sqrt{N}$–consistent and the asymptotic covariance matrix follows from standard considerations for nonlinear estimation.

The actual calculation of the Box–Cox quantile regression can be achieved in a simple way as follows. Chamberlain (1994) and Buchinsky (1995) suggest the following two step procedure based on the invariance property of quantiles with respect to a strictly positively monotonic transformation:

1. estimate $\beta^\theta(\lambda)$ conditional on λ by: $\quad \hat{\beta}_\theta(\lambda) = \text{argmin}_\beta \quad \sum_i \rho_\theta(T_{\lambda,i} - x_i'\beta)$,
2. estimate λ by solving: $\quad \min_{\lambda\in\mathcal{R}} \quad \sum_i \rho_\theta(T_i - (\lambda x_i'\hat{\beta}^\theta(\lambda) + 1)^{1/\lambda})$.

When implementing the two step procedure, Fitzenberger et al. (2004) encountered the following general numerical problem. For every λ, it is not guaranteed that for all observations $i = 1, ..., n$ the inverse Box-Cox transformation $\lambda x_i'\hat{\beta}_\theta(\lambda) + 1$ is strictly positive. However, this is necessary to implement the second step of Buchinsky's procedure. It is natural to omit the observations for which this condition is not satisfied. But this raises a number of problems. First, the set of observations omitted changes when going through an iterative procedure to find the optimal λ. Second, it is not a priori clear how such an omission of observations affects the asymptotic distribution of the resulting estimator. Third, should still the full set of observations be used in the first step?

The modification of the estimator suggested in Fitzenberger et al. (2004) consists of using only those observations in the second step for which the second stage of the estimation is always well defined for all λ in the finite interval $[\underline{\lambda}, \overline{\lambda}]$ and it is assumed that the true λ lies in this interval. The limits of the interval $\underline{\lambda}$ and $\overline{\lambda}$ are fixed ex ante. The set of admissible observations in the second step is chosen by estimating the first step above for both $\lambda = \underline{\lambda}$ and $\lambda = \overline{\lambda}\}$ and taking only those observations i for which $\lambda x_i' \hat{\beta}^{\theta}(\lambda) + 1 > 0$ for both values of λ. The first step is still implemented based on all observations which allows asymptotically for a more efficient estimator.

Fitzenberger et al. (2004) motivate the suggested modification by a theoretical result and an simulation study. Their Proposition 1 implies for the case of a bivariate quantile regression (one regressor plus an intercept) that if $\lambda x_i' \hat{\beta}_{\theta}(\lambda) + 1 > 0$ holds for both $\lambda = \underline{\lambda}$ and $\lambda = \overline{\lambda}\}$ then the inequality also holds for every $\lambda \in (\underline{\lambda}, \overline{\lambda})$. Even though this result does not hold for quantile regression with more than one regressor, violations only occur under extremely rare circumstances. The simulation results suggest that the modification works very well in general.

Box–Cox quantile regression can also be implemented in the case of right censoring by adding the censoring points to the minimization problem in the two steps of the procedure described above. Simulation experience (simulation results are available upon request) indicates that the suggested modification by Fitzenberger et al. (2004) works well for censored Box-Cox quantile regressions only up to an upper and lower bound of λ. These bounds seem to depend on the simulation design. Further research is necessary on this issue.

2.5 Unobserved Heterogeneity

In duration analysis, unobserved heterogeneity in the form of spell specific, time–invariant location shifts of the hazard rate or the duration distribution play a key role (Wooldridge, 2002, chapter 20.3.4, van den Berg, 2001). The popular mixed proportional hazard model (MPHM) assumes that the spell specific effect α enters the specification of the hazard rate in a multiplicative fashion $\lambda_i(t) = exp(x_i'\tilde{\beta})\lambda_0(t)exp(\alpha)$. Under the assumptions of the MPHM, α is a random effect which is distributed independently from the vector of covariates. It is well known that ignoring the presence of the random effect α will lead to misleading evidence on the shape of the baseline hazard $\lambda_0(t)$ inducing spurious duration dependence due to the sorting of spells with respect to α. Spells with a low value of α tend to survive relatively long and, thus, one might conclude that the hazard rate declines with elapsed duration when ignoring the influence of α. In general, ignoring the random effect α also biases the estimated coefficients for the covariates (Lancaster, 1990, p.65), though the impact is typically small. In the accelerated failure time model (equations 3 and 4), the random effect results in another component of the error term which is independent

of the covariates. Therefore, with known integrated baseline hazard ($\Lambda_0(.)$), quantile regression (or even OLS in the absence of censoring) can estimate consistently the coefficient estimates.

Quantile regression estimate the conditional quantile $q_\theta(T_i|x_i)$. Clearly, the increase in the conditional quantiles $q_\theta(T_i|x_i)$ for given x_i with increasing θ corresponds to the shape of the hazard rate as a function of elapsed duration. Thus, the increase in the conditional quantile is affected by the presence of unobserved heterogeneity. If the data generating process is an MPHM, then $\partial q_\theta(T_i|x_i)/\partial\theta$ differs from the average increase $E_\alpha\left\{\partial q_{\tilde{\theta}(\alpha)}(T_i|x_i,\alpha)/\partial\theta\right\}$ evaluated at the same duration level $q_\theta(T_i|x_i)$ corresponding to the quantile position $\tilde{\theta}(\alpha)$ for each α. This is due to the well known sorting effects in α ("low α" types tend to survive longer) and, therefore, the former term $\partial q_\theta(T_i|x_i)/\partial\theta$ is typically larger than the latter averaged version across α for small durations and smaller for larger durations. However, for an MPHM, the presence of a random effect typically causes only a small bias on the point estimates of the estimated quantile regression coefficients of the covariates because of the following argument.[10] Applying the implicit function to $S_i(q_\theta(T_i|C_i)|C_i) = 1-\theta$, both for $C_i = (x_i,\alpha)$ and $C_i = x_i$, results in

$$\frac{\partial q_\theta(T_i|C_i)}{\partial x_i} = -\left\{\frac{\partial S_i(t|C_i)}{\partial t}|_{t=q_\theta(T_i|C_i)}\right\}^{-1}\frac{\partial S_i(t|C_i)}{\partial x_i}|_{t=q_\theta(T_i|C_i)}\,. \tag{9}$$

Since $S(t|x_i,\alpha) = exp\{-exp(x_i'\tilde{\beta})exp(\alpha)\Lambda_0(t)\}$ and $S(t|x_i) = E_\alpha\{S(t|x_i,\alpha)\}$, it follows that

$$\frac{\partial q_\theta(T_i|C_i)}{\partial x_i} = \frac{-\tilde{\beta}\Lambda_0(t)}{\lambda_0(t)} \tag{10}$$

for $t = q_\theta(T_i|C_i)$ and $C_i = (x_i,\alpha)$ or $C_i = x_i$. This argument applies in an analogous way using a smooth monotonic transformation of the response variable. Thus, the estimated quantile regression coefficients only depend upon the coefficients $\tilde{\beta}$ and the shape of the baseline hazard. The quantile regression coefficients conditional upon α are the same for the same elapsed duration t irrespective of its rank θ, i.e. the estimated quantile regression coefficients for some unconditional quantile of the elapsed duration reflect the sensitivity of the respective quantile lying at this elapsed duration conditional upon the random effect. Put differently, some fixed duration $\bar{t}$ in general corresponds to two different ranks θ or $\theta(\alpha)$, respectively, when conditioning on $C_i = (x_i,\alpha)$ or $C_i = x_i$, i.e. $\bar{t} = q_\theta(T_i|x_i) = q_{\theta(\alpha)}(T_i|x_i,\alpha)$. Then, equation (10) implies that the partial effects for the different corresponding quantile regressions at this duration $\bar{t}$ are the same, i.e. $\partial q_\theta(T_i|x_i)/\partial x_i = \partial q_{\theta(\alpha)}(T_i|x_i,\alpha)/\partial x_i$. In this sense, a quantile regression on x_i provides meaningful estimates of partial effects, although the data are generated by an MPHM.

Evidence based on quantile regression can also be informative about the validity of the MPHM. Analogous to the PHM, a finding of crossings of conditional quantiles constitutes a rejection of

[10] See Zhang (2004) for detailed Monte Carlo evidence.

the MPHM. If $x_i'\tilde{\beta} < x_j'\tilde{\beta}$ for a pair (x_i, x_j), then the hazard is higher for x_j than for x_i for all α and therefore $S(t|x_i, \alpha) > S(t|x_j, \alpha)$, see line before equation (10). Integrating out the distribution of α, one obtains the inequality $S(t|x_i) > S(t|x_j)$ for all t. Thus, $q_\theta(T_i|x_i) > q_\theta(T_j|x_j)$ for all θ and therefore the MPHM implies that there should not be any crossings when just conditioning on the observed covariates. Intuitively, the independence between α and x_i implies that a change in covariates changes the hazard rate in the same direction for all α's. Therefore all quantiles conditional on (x_i, α) move into the opposite direction. The latter implies that the quantile conditional on only x_i must also move into that direction.

Instead of assuming that the random effect shifts the hazard rate by a constant factor as in the MPHM, a quantile regression with random effects for log durations could be specified as the following extension of the accelerated failure time model in equation (3)[11]

$$log(T_i) = x_i'\beta^\theta + \alpha + \epsilon_i^\theta \tag{11}$$

where the random effect α enters at all quantiles. The entire distribution of log durations is shifted horizontally by a constant α, i.e. $log(T_i) - \alpha$ exhibits the same distribution conditional on x_i. α is assumed independent of x_i and ϵ_i^θ. The latter is defined as $\epsilon_i^\theta = log(T_i) - q_\theta(log(T_i)|x_i, \alpha)$.[12] The regression coefficients β^θ now represent the partial effect of x_i also conditioning upon the random effect α. Such a quantile regression model with random effects has so far not been considered in the literature. It most likely requires strong identifying assumptions when applied to single spell data. Here, we use the model in (11) purely as point of reference.

How are the estimated quantile regression coefficients affected by the presence of α, when just conditioning on observed covariates x_i? Using $S(log(t)|x_i) = E_\alpha\{S(log(t)|x_i, \alpha)\}$ and result (9), it follows that[13]

$$\frac{\partial q_\theta(log(T_i)|x_i)}{\partial x_i} = \int \frac{f(\bar{t}\,|x_i)|x_i, \alpha)}{f(\bar{t}\,|x_i)|x_i)}\, \beta^{\tilde{\theta}(\alpha)}\, dG(\alpha) \tag{12}$$

for $\bar{t} = q_\theta(log(T_i)|x_i)$, where $f(.)$ and $F(.)$ are the pdf and the cumulative of the duration distribution, respectively, $G(.)$ is the distribution of α, and $\tilde{\theta}(\alpha) = F(q_\theta(log(T_i)|x_i)|x_i, \alpha)$. All expressions are evaluated at the duration $\bar{t}$ corresponding to the θ-quantile of the duration distribution conditioning only upon x_i. Hence, $f(q_\theta(log(T_i)|x_i)|x_i, \alpha)$ is the pdf conditional on both x_i and α and $f(q_\theta(log(T_i)|x_i)|x_i)$ the pdf just conditional on x_i both evaluated at the value of the quantile conditional on x_i. For the derivation of (12), note that $f(q_\theta(log(T_i)|x_i)|x_i) = \int f(q_\theta(log(T_i)|x_i)|x_i, \alpha)\, dG(\alpha)$.

[11] The following line of arguments applies analogously to the case with a general transformation $h(.)$.

[12] If ϵ_i^θ is independent of (x_i, α), then all coefficients, except for the intercept, can be estimated consistently by a quantile regression on just x_i. Also in this case, all slope coefficients are constant across quantiles.

[13] Only after submitting this paper, we found out that the result (12) is basically a special case of Theorem 2.1 in Hoderlein and Mammen (2005).

For the value $\bar{t} = q_\theta(log(T_i)|x_i)$, $\tilde{\theta}(\alpha)$ is the corresponding quantile position (rank) in the distribution conditioning both upon x_i and α. According to equation (12), the quantile regression coefficients conditioning only on x_i estimate in fact a weighted average of the β^θ in equation (11) where the weight is given by the density ratio for the duration $q_\theta(T_i|x_i)$ conditioning on both x_i and α and only on x_i, respectively. Since these weights integrate up to unity, the quantile regression estimate conditioning on x_i correspond to a weighted average of the true underlying coefficients in equation (11).

One can draw a number of interesting conclusions from the above result. First, if β^θ does not change with θ, then the estimated coefficients are valid estimators for the coefficients in equation (11). Second, if β^θ only change monotonically with θ, then the estimated coefficients will move in the same direction, in fact, understating the changes in β^θ. In this case, the random effect results in an attenuation bias regarding the quantile specific differences. Third, if one finds significant variation of the coefficients across quantiles, then this implies that the underlying coefficients in (11) exhibit an even stronger variation across quantiles. If the variation in the estimates follows a clear, smooth pattern, then it is most likely that the underlying coefficients in (11) exhibit the same pattern in an even stronger way.

Though being very popular in duration analysis, the assumption that the random effect and the covariates are independent, is not credible in many circumstances, for the same reasons as in linear panel data models (Wooldridge, 2002, chapters 10 and 11). However, fixed effects estimation does not appear feasible with single spell data. Identification is an issue here.

Summing up, though as an estimation method quantile regression with random effects have not yet been developed, it is clear that quantile regression conditioning just on the observed covariates yield meaningful results even in the random effects case. Changing coefficients across quantiles imply that such differences are also present in the underlying model conditional upon the random effect. Such a finding and the even stronger finding of crossing of predicted quantiles constitute a rejection of the mixed proportional hazard model, analogous to the case without random effects as discussed in section 2.1.

3 Unemployment Duration among Young Germans

The foregoing section gave an overview on the use of quantile regression for duration analysis. The purpose of this section is to apply censored quantile regression (CQR) to unemployment duration among German Youth in order to illustrate the applicability of the method. Youth unemployment is a critical issue in the literature. Our CQR results suggest that a simple proportional hazard rate model with time–invariant covariates is not justified for the data.

3.1 Data and Institutions

We use German register data for the analysis which contains spell information of employment and un-/nonemployment trajectories of about 500,000 individuals from West Germany. [14] More specifically, we use a sample of young unemployed workers that we drew from the IAB employment subsample (IABS) 1975–1997 (regional file).[15] IABS data have recently been used intensively for the analysis of unemployment duration. Plassmann (2002), Fitzenberger and Wilke (2004), and Biewen and Wilke (2005) investigate how the entitlement length for unemployment benefit affects the length of unemployment duration. Fahrmeier et al. (2003), Lüdemann et al. (2004), and Wilke (2005) investigate general determinants for the length of unemployment duration in West-Germany. Arntz (2005) analyzes the regional mobility of unemployed workers. The latter three papers use samples of unemployed aged 26–41. To our knowledge, no research of this type exists that primarily focuses on unemployment duration for young workers who are less than 26 years old.

The IAB employment subsample is a 1% sample of the socially insured working population. It has the general drawback that it does not contain periods of self-employment nor of employment as life-time civil servant (Beamte). The data provides daily information about the starting and the ending points of socially secured employment as well as unemployment provided that any form of unemployment compensation from the federal employment office (BA) is received. Receipt of social benefits or unemployment periods without the receipt of unemployment compensation are not recorded. Registered unemployment is therefore not directly observable and unemployment periods need to be constructed from the employment trajectories given the available information on income transfers, see Fitzenberger and Wilke (2004) for further details. In this paper, we adopt the "Nonemployment" proxy for our analysis, which can be considered as an upper bound of the true unemployment duration (Fitzenberger and Wilke, 2004):

- **Nonemployment** (NE): all periods of nonemployment after an employment period which contain at least one period with income transfers by the German federal labor office. The nonemployment period is considered as censored if the last record involves an unemployment compensation payment that is not followed by an employment spell.[16]

[14] In this analysis an individual is said to be West German if the last employment period before unemployment was in West Germany.

[15] For a general description of the data see Bender et al. (1996) and Bender et al. (2000). We complement our empirical analysis with some descriptive evidence for the year 1999 computed with the IAB employment subsample 1975-2001 - regional file -.

[16] A nonemployment spell is treated as right censored if it is not fully observed. All spells are censored at the end of 1997.

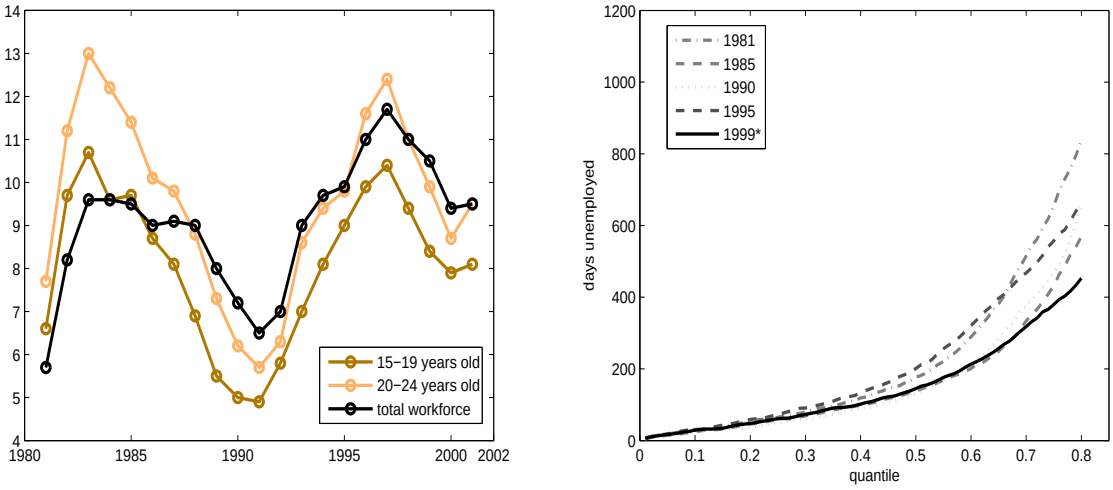
At least a part of each nonemployment period overlaps with unemployment and rules out purely out-of-the-labor-market periods. By construction it may contain periods of out of the labor market and it is therefore an upper bound of the true unemployment duration.

All unemployment periods in our sample require a foregoing employment period. This restriction causes a sample selection issue by ignoring all unemployment periods which do not meet this requirement. Note that the receipt of unemployment compensation and a foregoing employment spell are directly related. With the underlying data structure it is therefore hard to identify unemployment periods for young workers before their first job. Hence, we only consider unemployment after employment. The proportion of excluded unemployment periods is unknown. It is also not clear whether the results of this paper carry over to school leavers and related groups without any working experience. Note, that this restriction is not too binding since the majority of young workers start their working career by obtaining a vocational training degree through an apprenticeship with a firm which is recorded as employment in the data. Thus, our analysis covers the cases of young workers experiencing unemployment immediately after completion of the apprenticeship.

Analogous to Fitzenberger and Wilke (2004) and Wilke (2005) who analyze other age groups, we focus our analysis on four calender years of entry into Nonemployment to capture different macroeconomic conditions. Our sample consists of 23.705 spells. We include several demographic variables, work history variables, and regional information. In order to control for aggregate and regional labor market conditions, we include indicators for the calender year, quarterly GDP growth, and the monthly unemployment rate at the state level federal employment office districts (Landesarbeitsämter). In addition, we control for seasonal unemployment by adding a recall dummy and a winter dummy, both described below.[17] The evolution of group specific unemployment rates are shown in figure 1 (left graph). It is evident that the unemployment rate of young people in West-Germany decreased relative to the total unemployment rate. Unemployment rates for young people in 1997 (with the highest unemployment rate ever in Germany) were below the maximum numbers during the 1980s. The years 1981 and 1990 are similar in terms of the level of unemployment as well as 1985 and 1995. Figure 1 (right graph) presents nonparametric estimates of the quantile functions in the four years of interest and in 1999. In contrast to results on unemployment levels, the quantile functions for 1985 and 1990 are almost identical. Surprisingly, 1981 and 1995 are also similar when we ignore the highest quantiles indicating that longer durations are lower in 1995 compared to 1981.[18] Although the unemployment rate in 1999 was higher than

[17]Estimating a richer specification showed that age and the educational degree do not have explanatory degree for the length of unemployment periods. For this reason they are mostly not considered in the presented estimations.

[18]Note that the estimated quantile function of 1995 predicts less than 700 days at quantile 0.8 and is therefore



* estimate based on the IABS 1975-2001

Figure 1: Age group specific unemployment rates in West-Germany (Source: IAB Nuremberg) (left); nonparametric unconditional quantile functions for < 26 years old (right)

in 1995, the respective quantile function show a reversed different pattern. The quantile function in 1999 is even weakly below the function in 1990 during the economic boom period. Despite the descriptive nature of these findings, it is apparent that total unemployment is not a good predictor of unemployment durations for young workers. The following analysis uses therefore monthly regional unemployment rates which provide more succinct information.

Specifically, we include the following set of covariates for estimation purposes:

- gender of the unemployed.
- marital status of the unemployed.
- married females. This variable absorbs possible effects of out of the labor force periods or parental leave periods of females. A child indicator is not used because information about the presence of children is not available for all years.
- married females in the 1990s. This variable accounts for a possible change in the labor force participation rate of the females in the 1990s.
- we use 5 business sector categories: agriculture, trade/services/traffic, construction, public sector, others (mainly production). The variables are grouped according to the similarity of the results in preliminary estimations.

below the censoring of the data at the end of 1997.

- wage quintile 4,5. This variables indicates whether the unemployed had pre–unemployment earnings in the highest 40% of the population earnings distribution.
- quarterly GDP growth rate. This variable is the quarterly change in gross GDP measured in prices of 1995. It is affected by seasonal variation and by the business cycle in general.
- recall dummy indicating a recall to the former employer at the end of the previous unemployment period.
- four categories for regional information: Rhineland-Palatinate/Hesse, Baden-Wuerttemberg, Bavaria and northern states/Saarland (reference group). These categories are also constructed according to similarity in preliminary estimation results.
- monthly unemployment rate at the state level unemployment office districts which mainly overlap with the areas of the federal states.
- length of tenure before unemployment: less than one year, 1-3 years (reference category), more than 3 years.
- three calender year indicators: 1981 (reference year), 1985, 1990, 1995.
- indicator for no completed apprenticeship in 1995.
- winter time indicator, which is one if the unemployment spell starts between October and March.

Information about the educational degree and age variables are mostly not considered because preliminary estimates did not suggest that these variables have a sizeable effect. A descriptive summary of the sample is presented in table 1 in the appendix. Only about 8% of the unemployment spells are right censored, which is quite a moderate degree of censoring in the context of duration analysis.

3.2 Estimation Results

We estimate a log-linear version of model (1), i.e.

$$log(T_i) = x_i'\beta^\theta + \epsilon_i^\theta,$$

at each decile, $\theta = 0.1, 0.2, \ldots, 0.9$. The CQR model is estimated using the BRCENS algorithm implemented in TSP. Inference on the CQR coefficients is based on 500 bootstrap resamples using

the method of Bilias et al. (2000). The estimations involve the 22 covariates of table 1 and the constant. The estimated coefficients are reported in figures 2 and 3. It is evident that many of them vary significantly over the quantiles. The magnitude or even the sign of the effect of a covariate then differs between short-term (lower quantiles) and long-term unemployed (upper quantiles) indicating the need for a flexible estimation method. For example, the coefficient for the covariate "married" is not significant for short unemployment duration but its magnitude increases continuously and it is highly negative at the upper quantiles. The winter time variable elongates short duration and shortens long duration. However, the calender time effect has to be considered jointly with the quarterly GDP growth rate, the calender year dummies, and the monthly unemployment rate. For this reason, it is hard to interpret this result. As outlined in the previous section, the Cox-proportional hazard model does not allow for a change of sign of the partial derivative of the conditional quantile function, as observed very clearly for the coefficient of the winter dummy.

Let us now turn to a more detailed discussion of the estimation results and compare them to the results of Lüdemann et al. (2004) for middle aged workers using the same econometric model with different data. Unmarried females exhibit shorter unemployment spells than unmarried males. Married males have shorter unemployment duration than unmarried males. Married females have the longest unemployment duration and the difference to the other groups is very strong. This reflects periods out of the labor force coinciding with parental leave decisions. Unemployment periods of married females became shorter during the 1990s. These results are broadly in line with the findings of Lüdemann et al. (2004). A high level of earnings or a recall to the former employer in the past induce much shorter long term unemployment periods. The regional unemployment rate is not significant at any quantile of the distribution and underpins the findings of Lüdemann et al. (2004) that the unemployment rate does not have an explanatory degree for the length of unemployment periods in Germany. It is observed that unemployment periods in the South German federal states are shorter and in particular long term unemployment periods are much shorter in Bavaria than in the north of Germany. This is in line with the findings of Fahrmeir et al. (2003) and Arntz (2005) who use other samples of unemployed and different estimation techniques. The results for length of tenure suggests that unemployment duration are longer if the foregoing employment spell is short (<1 year) or long (>3 years). Turning to the calender time effect we find that unemployment periods are shorter in 1985 and 1990. Interestingly, they have almost the same length in 1981 and 1995, with the exception that periods of long term unemployment are much shorter in 1995.[19] This observation coincides with the shape of the

[19]The highest quantile in 1995 should be read with caution because estimation results may be affected by the systematic censoring at the end of the data in 1997.

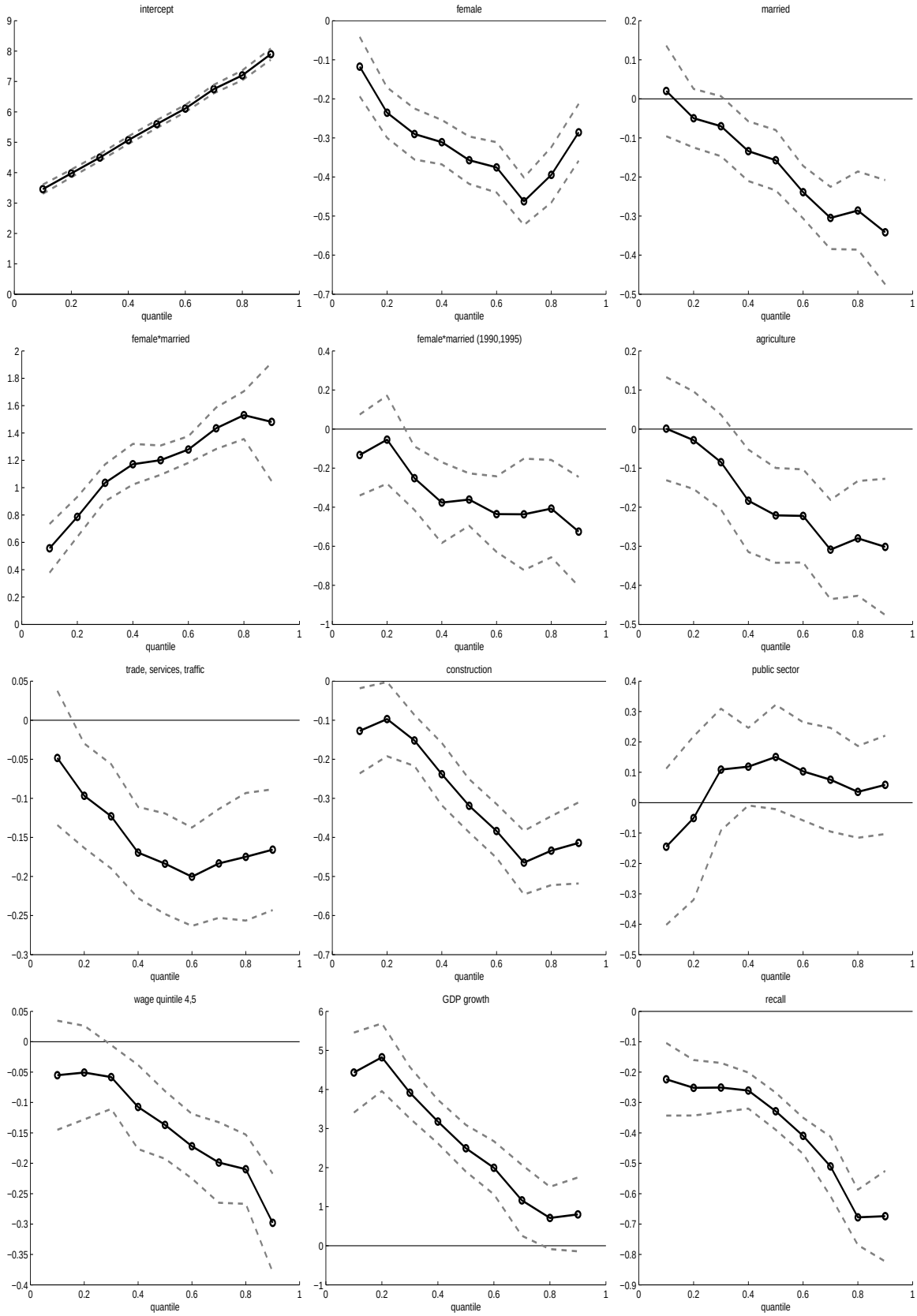


Figure 2: Estimated quantile regression coefficients with 90% bootstrap confidence bands, part I

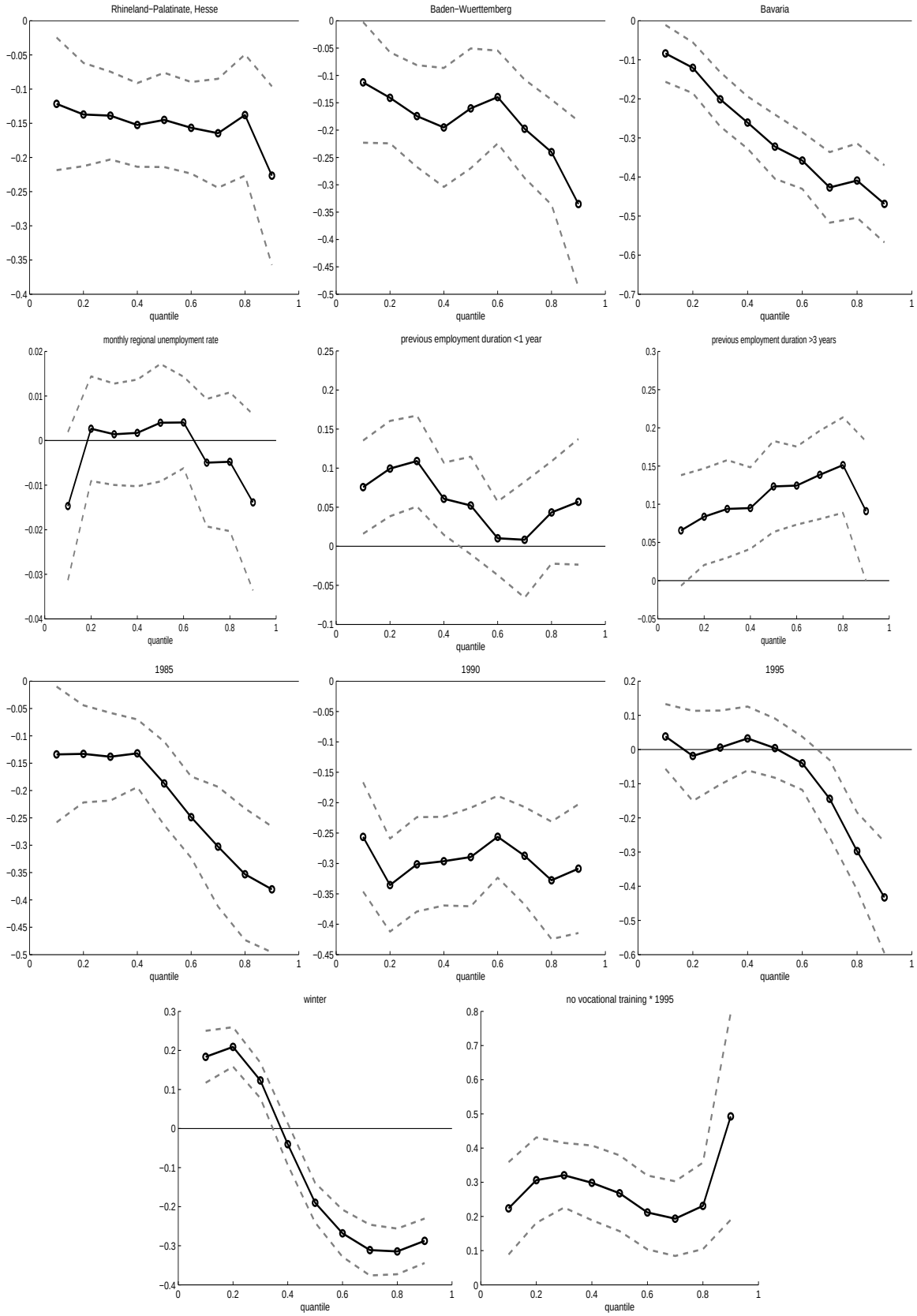


Figure 3: Estimated quantile regression coefficients with 90% bootstrap confidence bands, part II

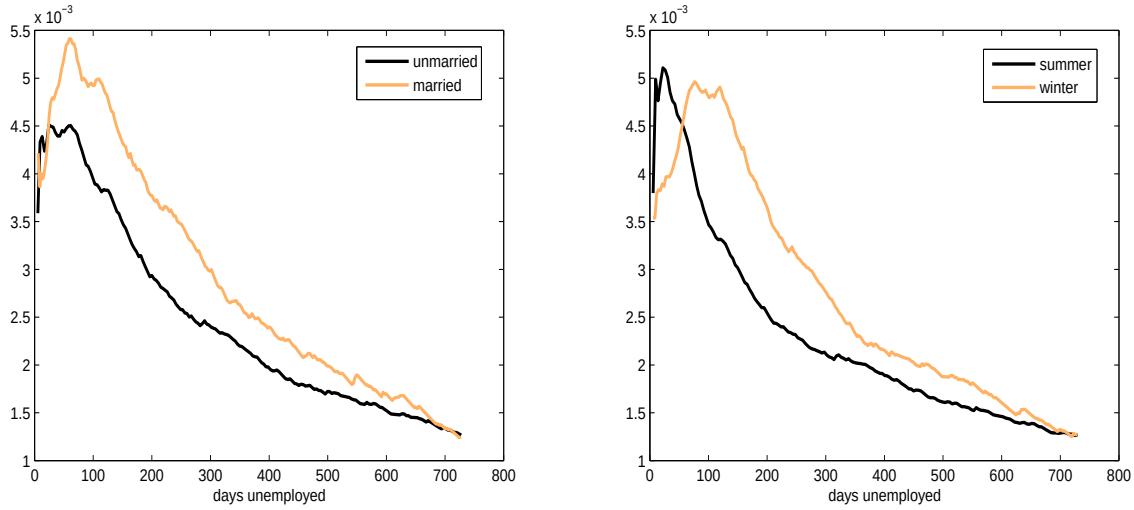


Figure 4: Estimated conditional hazard rates evaluated at sample means of the other regressors.

nonparametric quantile functions in figure 1, right. It is maybe related to policy measures against long-term unemployment of young workers that were possibly conducted during this period.

As shown in section 2.3 one can also estimate hazard functions using the GMP resampling procedure. Figure 4 presents different hazard rate estimates based on this methodology. For the reasons provided at the end of section 2.3 we base our hazard rate estimations on nonparametric density estimates for the log duration. The number of resamples is 500. Since almost 10% of our observations are right censored, we draw θ_m from the uniform distribution on $(\theta_l, \theta_u) = (0, 0.9)$. Both plots in figure 4 show that the estimated hazard rates are non–proportional over the duration time. The proportional hazard assumption is apparently violated in this application.[20]

4 Summary

This survey summarizes recent estimation approaches using quantile regression for (right-censored) duration data. We provide a discussion of the advantages and drawbacks of quantile regression in comparison to popular alternative methods such as the (mixed-)proportional hazard model or the accelerated failure time model. We argue that quantile regression methods are robust and flexible in a sense that they can capture a variety of effects at different quantiles of the duration distribution. Our theoretical considerations suggest that ignoring random effects is likely to have a smaller effect on quantile regression coefficients than on estimated hazard rates of proportional

[20]Note that we have not tested this statistically based on the estimated hazard rates. So far, no formal test procedure is available in the literature.

hazard models. Quantile regression do not impose a proportional effect of the covariates on the hazard. The proportional hazard model is rejected empirically when the estimated quantile regression coefficients change sign across quantiles and we show that this holds even in the presence of unobserved heterogeneity. However, in contrast to the proportional hazard model, quantile regression can not take account of time–varying covariates and it has not been extended so far to allow for unobserved heterogeneity and competing risks. We also discuss and slightly modify the simulation approach for the estimation of hazard rates based on quantile regression coefficients, which has been suggested recently by Machado and Portugal (2002) and Guimarães et al. (2004).

Our empirical application to unemployment duration data for young workers from West Germany demonstrates the usefulness of the discussed methods. Many estimated coefficients vary over the quantiles of the duration distribution and we observe changes of their sign. Competing alternative estimation approaches, such as a proportional hazard rate model with time–invariant covariates, cannot capture all these effects. Using data for workers aged 26–41, Lüdemann et al. (2004) observe similar violations of the proportional hazard assumption. Wilke (2005) observes crossings of the survivor functions. We conclude that the proportional hazard rate assumption is not justified in standard analysis of unemployment duration in Germany based on time–invariant covariates. Moreover, we present estimated conditional hazard rates based on quantile regression coefficients. These figures also suggest that the proportional hazard assumption is violated in our application. Our findings illustrate the usefulness of quantile regression for duration analysis.

References

Arntz, M. (2005). The geographical mobility of unemployed workers. Evidence from West-Germany. ZEW Discussion Paper 05-34.

Bender, S., Haas, A., and Klose, C. (2000). The IAB Employment Subsample 1975–1995. *Schmollers Jahrbuch*, Vol. 120, 649–662.

Bender, S., Hilzendegen, J., Rohwer, G., and Rudolph, H. (1996). Die IAB–Beschäftigtenstichprobe 1975–1990. Beiträge zur Arbeitsmarkt– und Berufsforschung. No. 197, Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (IAB) Nürnberg.

Biewen, M. and Wilke, R.A. (2005). Unemployment duration and the length of entitlement periods for unemployment benefits: do the IAB employment subsample and the German Socio-Economic Panel yield the same results? *Allgemeines Statistisches Archiv* 89(2), 209–236.

Bilias, Y., Chen, S., and Ying, Z. (2000). Simple Resampling Methods for Censored Regression Quantiles. *Journal of Econometrics*, 99, 373–386.

Box G. and Cox D. (1964). An Analysis of Transformation. *Journal of the Royal Statistical Society* B 26, 211–252.

Buchinsky, M. (1995). Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963-1987. *Journal of Econometrics* 65, 109–154.

Buchinsky, M. (1998). Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research. *Journal of Human Resources*, 33, 88–126.

Buchinsky, M. (2001) Quantile Regression with Sample Selection: Estimating Women's Return to education in the US. *Empirical Economics*, 26(1), 87–113.

Buchinsky, M. and Hahn, J. (1998). An Alternative Estimator for the Censored Quantile Regression Model. *Econometrica*, 66, 653–672.

Chamberlain, G. (1994) Quantile Regression, Censoring, and the Structure of Wages. In: Sims, C. (ed.), Advances in Econometrics: Sixth World Congress, Volume 1, Econometric Society Monograph.

Chen, S. (2002). Rank Estimation of Transformation Models. *Econometrica*, Vol. 70, No. 4, 1683–1697

Chernozhukov, V. and Hong, H. (2002). Three–Step Censored Quantile Regression and Extramarital Affairs. *Journal of the American Statistical Association*, 97, 872–882.

Fahrmeir, L., Lang, S., Wolff, J. and Bender, S. (2003): Semiparametric Bayesian Time-Space Analysis of Unemployment Duration. *Allgemeines Statistisches Archiv* 87, 281 – 307.

Fitzenberger, B. (1997). A Guide to Censored Quantile Regressions. In: *Handbook of Statistics, Volume 15: Robust Inference* (Eds. G.S. Maddala & C.R. Rao), 405-437. Amsterdam: North–Holland.

Fitzenberger, B. (1998) The Moving Blocks Bootstrap and Robust Inference for Linear Least Squares and Quantile Regressions. *Journal of Econometrics*, 82, 235–287.

Fitzenberger, B. and Wilke, R.A. (2004). Unemployment Durations in West-Germany Before and After the Reform of the Unemployment Compensation System during the 1980s. *ZEW Discussion Paper* 04-24.

Fitzenberger, B., Wilke, R.A. and Zhang, X. (2004). A Note on Implementing Box-Cox Regression. *ZEW Discussion Paper* 04-61.

Fitzenberger, B. and Winker, P. (2001). Improving the Computation of Censored Quantile Regressions. *Discussion Paper*, Universität Mannheim.

Guimarães, J., Machado, J.A.F. and Portugal, P. (2004). Has long become longer and short become shorter? Evidence from a censored quantile regression analysis of the changes in the distribution of U.S. unemployment duration. *Unpublished discussion paper*, Universidade Nova de Lisboa.

Hoderlein, S. and Mammen, E. (2005). Partial Identification and Nonparametric Estimation of Nonseparable, Nonmonotonous Functions. *Unpublished discussion paper*, University of Mannheim.

Horowitz, J. (1996). Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable . *Econometrica*, Vol. 64, No. 1, 103–137

Horowitz, J. and Neumann, G. (1989). Specification Testing in Censored Regression Models: Parametric and Semiparametric Methods. *Journal of Applied Econometrics*, 4, 61–86

Kiefer, N.M. (1988). Economic Duration Data and Hazard Functions. *Journal of Economic Literature*, XXVI: 649–679.

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

Koenker, R. and Bilias, Y. (2001). Quantile Regression for Duration Data: A Reappraisal of the Pennsylvania Reemployment Bonus Experiments. *Empirical Economics*, 26, 199–220

Koenker, R. and Geling, O. (2001). Reappraising Medfly Longevity: A Quantile Regression Survival Analysis. *Journal of the American Statistical Association*, 96(454), 458–468.

Koenker, R. and Hallock, K. (2002). Quantile Regression. *The Journal of Economic Perspectives*, 15(4), 143–156.

Lancaster, T. (1990). The Econometric Analysis of Transition Data. Econometric Society Monographs No. 17, Cambridge University Press.

Lüdemann, E., Wilke, R.A. and Zhang, X. (2004). Censored Quantile Regression and the Length of Unemployment Periods in West-Germany. ZEW Discussion Paper 04-57.

Machado, J.A.F. and Portugal, P. (2002). Exploring Transition Data through Quantile Regression Methods: An Application to U.S. Unemployment Duration. In: *Statistical data analysis based on the L1-norm and related methods – 4th International Conference on the L1-norm and Related Methods* (Ed. Yadolah Dodge), Basel: Birkhuser.

Plaßmann, G. (2002). Der Einfluss der Arbeitslosenversicherung auf die Arbeitslosigkeit in Deutschland. *Beiträge zur Arbeitsmarkt- und Berufsforschung* No. 255, Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit, Nürnberg.

Portnoy, S. (2003). Censored Regression Quantiles. *Journal of the American Statistical Association*, 98(464), 1001–1012.

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

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

Powell, J.L. (1991). Estimation of monotonic regression models under quantile restrictions. In: W.Barnett, J.Powell, and G.Tauchen, eds., Nonparametric and semiparametric methods in Econometrics, (Cambridge University Press, New York, NY) 357–384.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

Van den Berg, G.J. (2001). Durations Models: Specification, Identification and Multiple Durations. In: J.J. Heckman, E. Leamer, editors, *Handbook of Econometrics*, Volume 5, Elsevier, Amsterdam, 3381–3460.

Wooldridge, J.M. (2002). Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge, Massachusetts.

Wilke, R.A. (2005). New Estimates of the Risk and the Duration of Unemployment for West-Germany. *Schmollers Jahrbuch* 128(2), 207–237.

Zhang, X. (2004). Neuere Entwicklungen in der Analyse von Verweildauermodellen mit Quantilsregressionen als Alternative zum konventionellen Modell der proportionalen Hazardrate. Diploma Thesis. Mannheim University.

Appendix

Table 1: Descriptive Statistics

Variable	Mean	Median	Std.Dev.	Minimum	Maximum
Unemployment Duration (days)	411	151	688	1	6119
continuous variables					
quarterly gdp growth	1%	2%	0.035	−7%	4%
regional unemployment rate (in %)	7.8	7.4	2.78	2.8	14.2

Dummy Variable	= 1 if	Mean
Censored	yes	8%
Gender	female	41%
Marital status	married	16%
Education	unskilled∗1995	8%
Professional Group		
	agriculture	3%
	trade/food/servives	5%
	construction	16%
	public sector	3%
wage quintile 4,5		19%
tenure	<1 year	37%
	>3 years	23%
Recall	yes	9%
federal state	Rhineland-Palatinate, Hesse	14%
	Baden-Wuerttemberg	13%
	Bavaria	20%
calender time	winter period	52%
	1985	33%
	1990	19%
	1995	18%