Discussion Paper No. 05-67

Application of a Simple Nonparametric Conditional Quantile Function Estimator in Unemployment Duration Analysis

Laura Wichert and Ralf A. Wilke

ZEW

Zentrum für Europäische
Wirtschaftsforschung GmbH

Centre for European
Economic Research

Discussion Paper No. 05-67

Application of a Simple Nonparametric Conditional Quantile Function Estimator in Unemployment Duration Analysis

Laura Wichert and Ralf A. Wilke

Non technical summary

Applied researcher often have to handle large data sets, e.g. administrative data has recently gained popularity in research on unemployment. In order to explore those data sets and to get a first impression about possible dependencies it is convenient to impose as few assumptions as possible about the structure of the econometric model.

In this paper, we present a nonparametric conditional quantile function estimator with mild functional assumption that can be used as a fast and powerful tool for data exploration. We apply the estimator to a sample of the "IAB-Beschäftigtenstichprobe" (IAB-employment sample) that includes daily employment trajectories of the socially insured workforce in Germany. The focus of our analysis is the impact of age and the previous wage level on the length of unemployment for short-, middle-, and long-term unemployed.

We find that age has a strong extending influence on unemployment of long-term unemployed, i.e. old long-term unemployed have much longer unemployment spells than younger unemployed. In the group of the short-term unemployed age doesn't matter: it takes a young short-time unemployed as long as an old one to find a job. This age pattern is much clearer for men than for women, which is likely linked to maternity leave of women.

As to the wage level before the beginning of unemployment, we find a weak negative, i.e. shortening, influence for short-term unemployed. This influence strengthens for the middle- and long-term unemployed up to a previous wage level of 65 Euro per day: In this group, persons who earned more before they became unemployed are shorter unemployed than those with a lower former income. In the group of long-term unemployed with a previous wage level of more than 80 Euro, we find a weak extending impact of a higher wage level. Especially the results for the long-term unemployed at a low previous wage level give an impression about the impact of the system of social benefits on the length of unemployment, since the financial incentive to leave unemployment is rather small for this group.

We provide a theoretical motivation for the estimator and show with simulations that it is fast and reliable in typical data structures.

Application of a Simple Nonparametric Conditional Quantile Function Estimator in Unemployment Duration Analysis.*

Laura Wichert†

Ralf A. Wilke‡

09.2005

*We thank Eva Müller for research assistance and Hidehiko Ichimura and Noël Veraverbeke for remarks on the paper. Comments from the seminar participants at Mannheim University and at the IAB Nutzertagung in Nuremberg are gratefully acknowledged. The authors gratefully acknowledge financial support by the German Research Foundation (DFG) through the research project *Microeconometric modelling of unemployment duration under consideration of the macroeconomic situation*. This work uses 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 Western Germany and from 1992 to 2001 in Eastern 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-Leistungsempfaenger-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.

†ZEW Mannheim and TU Darmstadt, laura.wichert@web.de

‡Centre for European Economic Research (ZEW Mannheim), P.O.Box 10 34 43, 68034 Mannheim, Germany, E-mail: wilke@zew.de

Abstract

In many econometric applications it is unclear from the very beginning whether a parametric functional of a continuous regressor should be specified as a linear, as a higher order polynomial or as a piecewise linear. Nonparametric estimators can provide relevant information as they are a convenient tool for data exploration. For this purpose we consider an extension of the conventional univariate Kaplan-Meier estimator for the hazard rate to multivariate right censored duration data and truncation of the marginal distributions of random regressors. It is a combination of nearest neighbor estimator and the Nelson-Aalen type estimator. It is a Akritas (1994) type estimator which adapts the nonparametric conditional hazard rate estimator of Beran (1981) to more typical data situations in applied analysis. We show with simulations that the estimator has nice finite sample properties and our implementation appears to be very fast. A small application to German unemployment duration data demonstrates the need for flexible specifications of conditional quantile functions. The results indicate that the level of social benefits has a strong impact on the length of long term unemployment for low earners and fertility is related with longer unemployment periods of females.
Keywords: unemployment duration, nonparametric conditional hazard rate, censoring, truncation

JEL: C14, C34, C41

1 Motivation

More and more national governments make administrative data available to the research community. These data sets are large, since they cover the full population. Applied researcher may therefore use flexible econometric frameworks with mild functional form assumptions for a detailed exploration of the data. In this paper we focus on nonparametric conditional quantile functions for right censored responses and with truncated marginal distributions of random regressors. Classical duration models, such as the accelerated failure time or the Cox model, require strong conditions (Koenker and Geling, 2001, Portnoy, 2004) that may not be met by the underlying empirical problem (Portnoy, 2004, Fitzenberger and Wilke, 2006). For this reason quantile regression is emerging as a popular alternative in applied economics, see Koenker and Bilias (2001), Machado and Portugal (2002), and others. However, linear or non linear quantile regression applied to right censored duration data (Fitzenberger and Wilke, 2005) may depend on x in a variety of ways and it is difficult in an application to explore what functional relationship may be appropriate. In addition, due to the nature of the data generation process the distribution of x may be truncated. The most common example in administrative data is an individual's wage, which is typically not observed below and above a certain limit.

Using the spirit of Beran (1981) we consider a nonparametric estimator for smooth α quantile functions $q_\alpha(x)$ if the distribution of x is truncated. We do not impose shape restrictions on the conditional density of the response y and the resulting estimate for $q_\alpha(x)$ can be used as an explorative tool. It may detect important nonlinearities in x and it may provide evidence whether the shape of the functional has restrictions across quantiles, e.g. $q_\alpha(x) = g(\alpha, q(x))$, where g is a known parametric functional with potentially unknown parameters. It can be used as a tool for more appropriate specification of a microeconometric model with more structure. We follow the nonparametric conditional hazard rate estimator of Beran (1981) with the main difference that we use a nearest neighbor estimator (Yang, 1981) for the smoothing in x. For this reason our estimator is also applicable when the distribution of regressors is truncated, which is as already mentioned a common problem in administrative data. Akritas (1994) considers a similar estimation strategy and he derives asymptotic properties for this class of estimators. Like in his and in Beran's pioneering work, we do not impose a smoothing in the dimension of the response as it is analyzed by McKeague and Utikal (1990) and van Keilegom and Veraverbeke (2001). This implies that our estimator is a step function in y. If y is continuous in an application one may use simple uniform weights in a neighborhood of the grid points on the support of the response.

The paper aims to convince the applied researcher that our estimation strategy is an applicable solution to common empirical problems for example in duration analysis. Finite sample performance and computing time are both promising. A small application to German administrative unemployment duration data shows that quantile functions may be nonlinear in the regressor and posses different shapes for different quantiles. This highlights the need for flexible specifications.

2 The Estimator

We consider a model with an unknown joint distribution (Y, X) which is smooth in X. Y is discrete response or duration and X are continuous regressors with truncated distributions. Suppose we have a random sample $(\tau_i, X_i, d_i)_{i=1,\ldots,n}$, where d_i is an indicator for censoring of Y_i.[1] Our aim is to estimate the smooth nonparametric conditional quantile function $q_\alpha(x)$ of the distribution of y conditional to x, $F(y|x)$, for right censored data:

$$q_\alpha(x) = \inf\{y|S(y|x) \geq \alpha\},$$

where $S(y|x) = 1 - F(y|x)$ is the conditional survivor function.

Let $\hat{h}(y)$ be the unconditional hazard rate of the distribution of y estimated with the classical Kaplan-Meier type estimator

$$\hat{h}(y) = \frac{\sum_{i=1}^{n} 1_{\tau_i = y} 1_{d_i=1}}{\sum_{i=1}^{n} 1_{\tau_i \geq y}},$$

[1] $d_i = 0$ if Y_i is censored and $\tau_i = min(Y_i, C_i)$ with C_i as upper threshold. Y_i and C_i are mutually independent given x.

where 1 is the indicator function. The numerator divided by n estimates the conditional probability $P(Y = y|d = 1)$[2] and the denominator divided by n estimates the survivor function of the response $P(Y \geq y)$.

In order to study regression problems with censored data, Beran (1981) suggests the so called conditional Kaplan-Meier estimator:

$$F(y|x) = \prod_{\tau_i \leq y} \left(1 - \frac{w_{n(i)}(x; b_n)}{1 - \sum_{j=1}^{i-1} w_{n(j)}(x; b_n)}\right)^{d_{(i)}};$$

where $w_{n(1)}(x; h_n) \leq \ldots \leq w_{n(n)}(x; h_n)$ are appropriate positive weights.[3] Beran assumes for simplicity ordered design points x on $[0, 1]$. In case of no censoring his estimator is equivalent to Stone's (1977) estimator. In the case of uniform weights $1/n$ it is the univariate Kaplan-Meier estimator or the Nelso-Aalen type estimator. In order to keep things simple we use simple Kernel weights instead of the Gasser-Müller weights in this paper.

In empirical situations the marginal distribution of the regressors may be truncated. For this reason we adopt in our approach to estimation the nearest neighbor design of Yang's (1981) SNN estimator. For simplicity let us first focus on a univariate x. Let $f(x)$ be the marginal distribution function of x. The SNN estimator

[2] If the distribution of Y is continuous it may be useful for finite sample reasons to use uniform weights in the neighborhood of y, i.e. as numerator $\sum_{i=1}^{n} 1_{\tau_i \in [y-\Delta, y+\Delta]} 1_{d_i=1}$ and as denominator $\sum_{i=1}^{n} 1_{\tau_i \geq y-\Delta}$ with $\Delta > 0$ and the ordered grid points y_j satisfy $y_j - y_{j-1} - 2\Delta = 0+$. This is in fact rounding of Y_i towards the closest grid point. Alternatively, one may use kernel smoothing in the dimension of Y as done by e.g. McKeague and Utikal (1990) and van Keilegom and Veraverbeke (2001).

[3] Beran (1981) and several subsequent papers use the Gasser-Müller weights which are defined as

$$w_{n(i)}(x; b_n) = \frac{1}{c_n(x; b_n)} \int_{x_{i-1}}^{x_i} \frac{1}{h_n} K\left(\frac{x - z}{b_n}\right) dz \qquad i = 1, \ldots, n,$$

$$c_n(x; b_n) = \int_0^{x_n} \frac{1}{b_n} K\left(\frac{x - z}{b_n}\right) dz,$$

where $x_0 = 0$ and K is the kernel function with b_n as the bandwidth sequence (sequence of positive constants, tending to 0 as n $\to \infty$).

is then given by:

$$f_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K\Big(\frac{G_n(x) - G_n(X_i)}{b_n}\Big),$$

where $G_n(x) = \sum_{i=1}^{n} \frac{1_{X_i \leq x}}{n}$ is the empirical distribution function of x. This estimator uses $G_n(x)$ instead of x. K is a continuous density function.[4] Using SNN kernel weights, we suggest the following estimator for the conditional hazard rate of y given x:

$$\hat{h}(y|x) = \frac{\sum_{i=1}^{n} 1_{\tau_i = y} 1_{d_i=1} K\left(\frac{G_n(x)-G_n(X_i)}{b_n}\right)}{\sum_{i=1}^{n} 1_{\tau_i \geq y} K\left(\frac{G_n(x)-G_n(X_i)}{b_n}\right)},$$

where the numerator and the denominator are smoothed estimates of the conditional probabilities $P(Y = y|d = 1, X = x)$ and $P(Y \geq y|X = x)$.[5] The advantages and disadvantages of Yang's estimator carry over to the estimation of conditional hazard rates: it is enough to have information about the rank of the X_i, since we do not require knowledge about its actual value. This implies that the estimator also works in the case of a truncated regressor. The SNN smoothing works like a variable bandwidth. Note that in case of arbitrary uniform weights K the estimator becomes the conventional Kaplan-Meier estimator.

According to Kaplan and Meier (1958), based on the hazard rate one can estimate the univariate survivor function for homogeneous right censored data with the product limit estimator:

$$\hat{S}(y) = \prod_{\tau_j < y} (1 - \hat{h}_j),$$

for $t \geq 0$, where the subscript j runs over the N grid points on the support of y.

Applying this procedure to our multivariate framework provides the conditional survivor function

$$\hat{S}(y|x) = \prod_{\tau_j < y} (1 - \hat{h}_j(y|x)),$$

[4] Yang (1981) requires several properties for K and the the choice of h in order to show mean squared and uniform convergence, see his theorems 2 and 3.

[5] It can be shown that the estimator of Beran (1981) is equivalent if the x values are ordered on their support $[0, 1]$ and if one would use Gasser-Müller weights. However, our implementation of the estimator does not require distinct values Y_i.

for $t \geq 0$ for an appropriate grid $j = 1, \ldots, N$ on the support of y.

The conditional quantile function at the αth quantile is now estimated by

$$\hat{q}_\alpha(x) = \inf\{y|\hat{S}(y|x) \geq \alpha\}.$$

An extension to multivariate x of dimension $k = 1, \ldots, m$ with $(\tau_i, X_i, d_i)_{i=1,\ldots,n}$ is straightforward by applying the idea of product kernels. In this case the conditional hazard is simply

$$\hat{h}(y|x) = \frac{\sum_{i=1}^{n} 1_{\tau_i = y} 1_{d_i=1} \prod_k K\left(\frac{G_n(x_k) - G_n(X_{ik})}{b_{nk}}\right)}{\sum_{i=1}^{n} 1_{\tau_i \geq y} \prod_k K\left(\frac{G_n(x_k) - G_n(X_{ik})}{b_{nk}}\right)}$$

and the remainder of the estimation strategy is unchanged. Note that this estimation strategy suffers from the curse of dimensionality and should not be applied if k is too large.

Akritas (1994) derives asymptotic properties for the estimator of the conditional survivor function. Weak convergence can be established by an appropriate choice of the bandwidth and under some technical assumptions. The numerator and the denominator of our estimator then converge to the conditional probabilities $P(Y = y|d = 1, X = x)$ and $P(Y \geq y|X = x)$ respectively. Akritas (1994) also derives an expression for the covariance function but alternatively he suggests the bootstrap method. We follow the second approach because it appears to be more simple.

3 Simulation

We perform simulation studies for two different model designs. For our first simulation we generate normally distributed random numbers $X \sim$N(5,1) and a random error term $\epsilon \sim$ N(0,0.5). In order to show the properties of the estimator for truncation, we truncate the distribution of x on both sides.[6] Let

$$y = x + \epsilon$$

[6] $X_i = 0$ if $X_i < 3$ and $X_i = 10$ if $X_i > 7$.

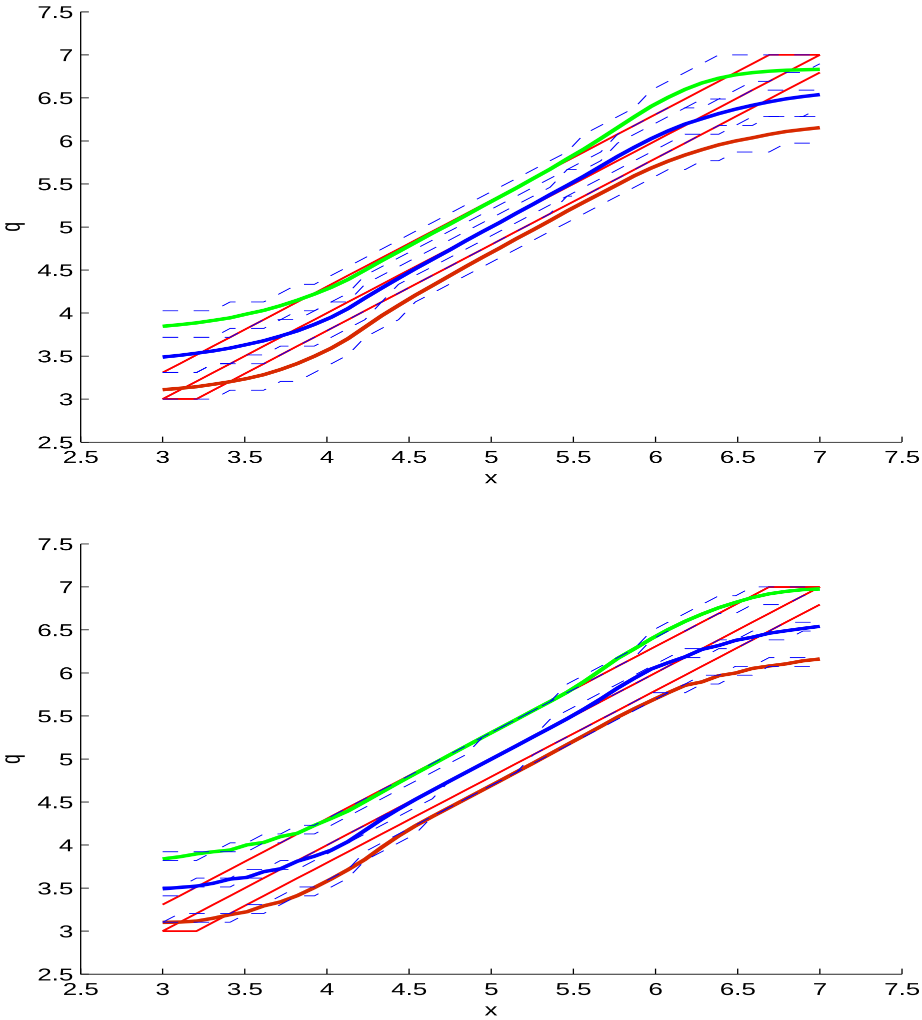

Figure 1: Simulation with 500 observations (**above**) and 5,000 observations (**below**); mean of the estimates of the quantile functions for α=0.3;0.5;0.7, the 2.5%- and the 97.5%-bootstrap quantile and the real model.

with an average right censoring of 10% for Y_i.[7] We simulate the 0.3, 0.5 and 0.7 quantile function 500 times for 500 observations and for 5,000 observations and calculate the mean of the simulations. As kernel function we use the Epanechnikov kernel function

$$K(x) = max(0.75 * (1 - x^2); 0).$$

For each quantile function, we also report the 2.5%- and the 97.5%-quantile of the sample of estimation results and the true quantile functions (see figure 1). The bandwidth in the kernel function is $b_n = 0.2$. The mean runtime for one simulation is 0.3 seconds for 500 observations and 2.5 seconds for 5,000 observations (AMD64 1.4 GHz, 64 Bit Linux, 64BIT Matlab v7.01) where we have 40 grid points on the support of x and 40 grid points on the support of y. This is evidently very fast.

The biased shape on both ends of the figure is due to the fact that our estimator fits locally a constant. Therefore we have a boundary bias that starts at a distance of the bandwidth apart of the edge of observations. Since we use the SNN smoothing we have a variable bandwidth given x. The low density of x at the boundaries implies a larger bandwidth than it would be without the SNN method. This leads to a stronger boundary bias in this case.

In order to check our estimator for nonlinear functions we simulate a second model:

$$y = 0.1 * x + 0.1 * x^2 - 0.1 * x^3 + \epsilon$$

again with an average right censoring of 10% for Y_i and a truncated distribution of x.[8] As before, we simulate the model 500 times, once with 500 observations and once with 5,000 observations and calculate the means. We also compute the 2.5%- and the 97,.5%-quantile of the resulting estimates distribution and the true quantile functions (see figure 2). The bandwidth of the kernel function is again $b_n = 0.2$. The runtime per simulation for the second model is about 0.5 second for 500 observations and 5 seconds for 5,000 observations.

[7] Y_i is considered censored if $Y_i > C_i$, with $C_i \sim$N(6.5,0.35).

[8] Same censoring as before, this time with $C_i \sim$N(-3.3,0.75) and $X_i = 0$ if $X_i < 4$ and $X_i = 10$ if $X_i > 6$.

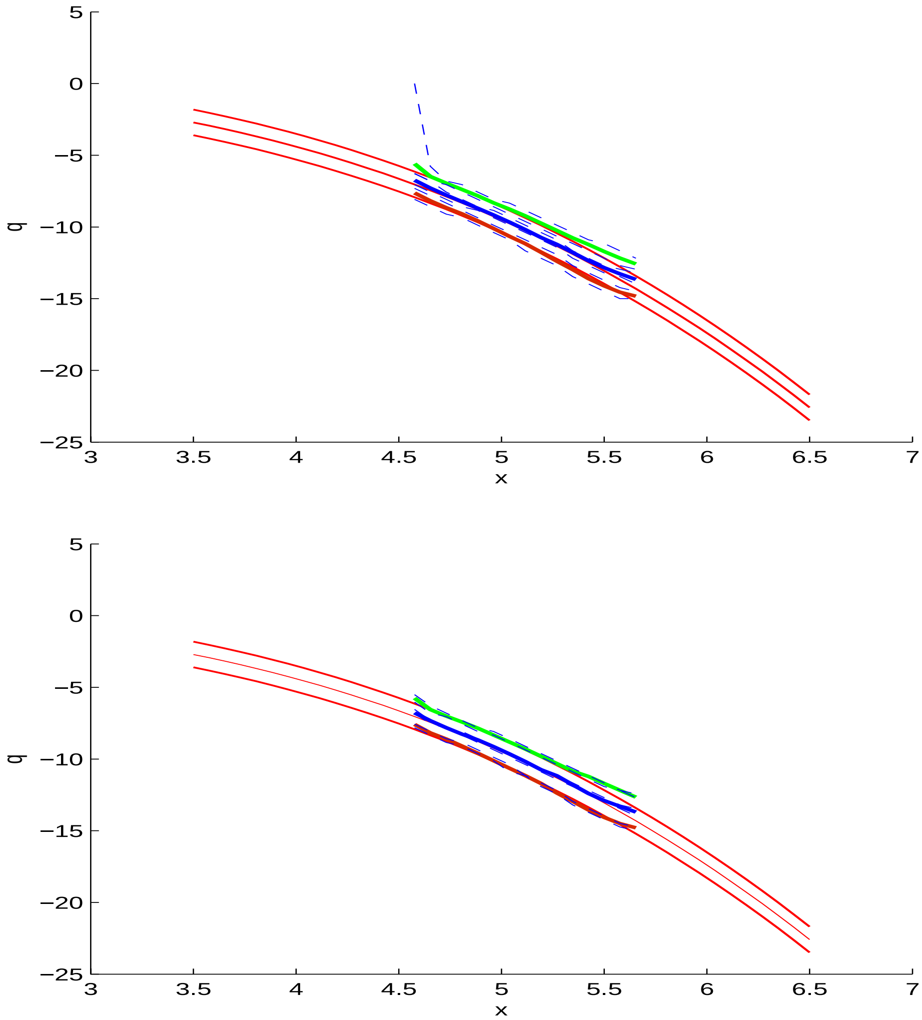

Figure 2: Simulation with 500 observations (**above**) and 5,000 observations (**below**); mean of the estimates of the quantile functions for α=0.3;0.5;0.7, the 2.5%- and the 97.5%-bootstrap quantile and the real model.

4 Empirical Results

We apply our estimator to a sample of the "IAB-Beschäftigtenstichprobe" (IAB-employment sample) 1975-2001 (IABS-R01) which contains daily employment trajectories of about 1.1 Mio indivduals from West-Germany and about 200K individuals from East-Germany. It is a 2% random sample of the socially insured workforce. In particular we use exactly the same sample of observations as is used by Fitzenberger and Wilke (2005). However, we restrict attention to the age, the gender and the last daily wage before unemployment for all "nonemployment"[9] spells starting in 1996 or 1997 in West-Germany. Our sample comprises 19,473 observations.[10]

Using this data we estimate the smooth nonparametric conditional quantile function of the distribution of unemployment duration conditional to age and to the previous wage level. For the estimation of the standard errors we use the bootstrap method by drawing 500 resamples with replacement and plot the 5- and the 95%-quantiles of the bootstrap distribution for the conditional quantile functions.

Effect of age Figure 3 shows the estimation results for the distribution of the length of unemployment conditional to age for the 0.3-, 0.5- and the 0.7-quantile for males (left) and females (right) with the 5- and 95%-bootstrap quantiles for each quantile. Comparing these results with other studies on this topic using quantile regression (Fitzenberger and Wilke, 2005), they don't leed to contradictions: While age plays a less important role for the shortly unemployed men and women, there is a strongly positive influence of age in the group of the long-term unemployed men. The pattern for the longer unemployed women isn't as clear as it is for men, especially not for the 0.7-quantile. According to Lechner [11] the probability of fertility has its maximum between the age of 26 and 30. This fact could offer a possible explanation

[9]We use the nonemployment proxy for unemployment introduced by Fitzenberger and Wilke (2004).

[10]For all relevant details on the sample used here see Fitzenberger and Wilke (2005).

[11]See Lechner (1997), p.6.

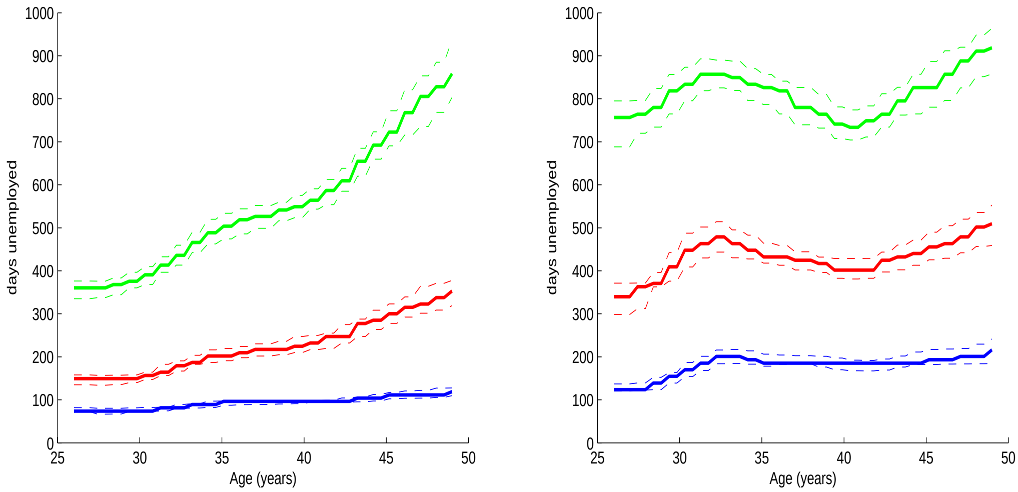

Figure 3: Estimated quantile functions conditional on age (for $\alpha = 0.3; 0.5; 0.7$); **left:** males, **right:** females

for the peak of the curve at the age of 32: At that age, mothers have passed their maternity leave and are considered unemployed but maybe not actually looking for a job. The obtained conditional quantile functions also support the findings of Biewen and Wilke (2005) who use semiparametric hazard rate and accelerated failure time models for both, males and females, but they provide more detailed insights.

Effect of previous wage Figure 4 shows the distribution of the length of unemployment spells conditional to the previous wage level, again for the 0.3-, 0.5- and the 0.7-quantile with the bootstrap quantiles. Because of some lack of information about part-time work which is rather frequent for females, we only regard the males in this case. As for the previous wage level, we only look at the span between the 10- and the 90%-quantile of former income. We only find a weak negative dependence of the length of unemployment on the previous wage level. At the 0.5-quantile, this dependence strengthens, especially until a previous wage level of 65 Euro per day. Only for the long-term unemployed with a previous wage level of more than 80 Euro per day we find small evidence that a high level of former income is linked with a shorter unemployment spell. As can be seen from the confidence intervals, the num-

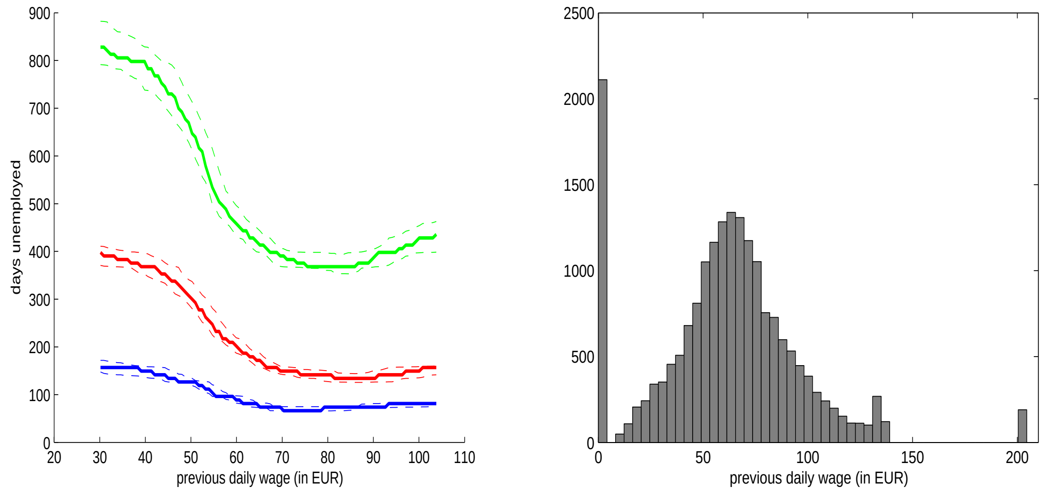

Figure 4: **left:** Estimated quantile functions conditional on the previous wage (for $\alpha = 0.3; 0.5; 0.7$) for males; **right:** Histogram of the previous wage for males

ber of observations of long-term unemployed coming from a well-payed job is limited, therefore the results have to be interpreted with caution. The long long-term unemployment periods for low wage individuals are probably related with the system of social benefits. The income transfers for this group do not decrease after expiration of unemployment benefits and moreover, it may be difficult for this group to find a job with payments significantly above the level of social benefits. The histogram in figure 4 shows the truncated distribution of the variable "previous wage" for males, whereas the value "0" means an income below and the value "200" means an income above the social security contribution ceiling ("Beitragsbemessungsgrenze").

5 Conclusion and Outlook

This note suggests a simple estimation strategy that is applicable to typical situations in applied economic research in particular in duration analysis when the distribution of the regressors is truncated. Our simulations and our application show that it is a fast and powerful tool for data exploration that works without strong assumptions. Results can be used for a more appropriate specification of a microeconometric model of more structure. There are several interesting topics for future research that may be beneficial for applied analysis: one could introduce a partially linear approach or allow for discrete regressors or an additive nonparametric structure. In our application we found some evidence that the conditional quantile functions possess different shapes across quantiles. Therefore one may also develop a test for shape invariance of those functions. Such a test would then provide elaborate information whether a more structural model, such as censored quantile regression, would require different model specifications across the quantiles.

References

Akritas, M.G. (1994), "Nearest neighbor estimation of a bivariate distribution under random censoring", *Annals of Statistics*, 22, 1299–1327.

Beran, R. (1981), "Nonparametric Regression with Randomly Censored Survival Data", Technical Report, University of California, Berkeley, CA.

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), 409–425.

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

Fitzenberger, B. and Wilke, R.A. (2005), "Conditional Hazards of Leaving Unemployment: an Application of Censored Box-Cox Quantile Regression to Administrative Data from Germany", *Mimeo*, ZEW Mannheim.

Fitzenberger, B. and Wilke, R.A. (2006), "Using Quantile Regression for Duration Analysis", *Allgemeines Statistisches Archiv*, 90(1), forthcoming.

Kaplan, E. L., and P. Meier (1958), "Nonparametric estimation from incomplete observations", *Journal of the American Statistical Association*, 53, 457–481.

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

Koenker, R. and Geling, O. (2001), "Reappraising Medfly Longevity: A Quantile Regression Survival Analysis", *Journal of the American Statistical Association*, Vol.96, No.454, 458–468.

Lechner, M. (1997), "Eine empirische Analyse der Geburtenentwicklung in den neuen Bundesländern", Beiträge zur angewandten Wirtschaftsforschung, No. 551-97.

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.

McKeague, I.W. and Utikal K.J. (1990), "Inference for a Nonlinear Counting Process Regression Model", *Annals of Statistics*, 18, 1172–1185.

Portnoy, S. (2004), "Censored Regression Quantiles", *Journal of the American Statistical Association*, Vol.98, No.464, 1001–1012.

Stone, C.J. (1977), "Consistent nonparametric regression", *Annals of Statistics*, 5, 595–645.

van Keilegom, I. and Veraverbeke, N. (2001), "Hazard Rate Estimation in Nonparametric Regression with Censored Data", *Ann. Inst. Statist. Math.*, 53, 730–745.

Yang, S. (1981), "Linear functions of concomitants of order statistics with applications to nonparametric estimation of a regression function", *Journal of the American Statistical Association*, 76, 658–662.