Nonparametric Kernel Estimation Methods for Discrete Conditional Functions in Econometrics

A THESIS
SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)
IN THE FACULTY OF HUMANITIES

2013

OBBEY AHMED ELAMIN

SCHOOL OF SOCIAL SCIENCE/ECONOMICS

Contents

List of Tables

List of Figures

Abbreviations

AFT	accelerated failure time
AMISE	asymptotic mean integrated squared errors
aMSE	average mean square error
BKH	Beran kernel hazard
BNL	binomial logit
CDF	conditional distribution function
CDNKH	conditional density-implied kernel hazard
CEKH	conditional external kernel hazard
CV	cross validation
DNKH	density-implied kernel hazard
DCF	discrete conditional function
EL	empirical likelihood
HPC	high performance computer
IAMSE	integrated average mean squared errors
IMSE	integrated mean squared error
ISE	integrated squared errors
KH	kernel hazard
KMWFB	kernel method with fixed bandwidths
LOO	leave-on-out
LSCV	least squares cross validation
ML	maximum likelihood
MLCV	maximum likelihood cross validation
MLCV-Risk	maximum likelihood cross validation risk
MSE	mean squared error
MDT	mixed data types
MPH	mixed proportional hazard
MNL	multinomial logit
MNP	multinomial probit
NPML	nonparametric maximum likelihood
PDF	probability density function
CDF	probability distribution function
PH	proportional hazard
RMT	random matching theory
RUT	random utility theory
EKH	external kernel hazard (the unconditional external kernel hazard)
UH	unobserved heterogeneity
WIMSE	weighted integrated mean squared error
WMSE	weighted mean squared errors

The University of Manchester
Obbey Ahmed Elamin
Doctorate of Philosophy in Economics
Nonparametric Kernel Estimation Methods for Discrete Conditional Functions in Econometrics
2013

Abstract

This thesis studies the mixed data types kernel estimation framework for the models of discrete dependent variables, which are known as kernel discrete conditional functions. The conventional parametric multinomial logit MNL model is compared with the mixed data types kernel conditional density estimator in Chapter (2). A new kernel estimator for discrete time single state hazard models is developed in Chapter (3), and named as the discrete time "external kernel hazard" estimator. The discrete time (mixed) proportional hazard estimators are then compared with the discrete time external kernel hazard estimator empirically in Chapter (4). The work in Chapter (2) attempts to estimate a labour force participation decision model using a cross-section data from the UK labour force survey in 2007. The work in Chapter (4) estimates a hazard rate for job-vacancies in weeks, using data from Lancashire Careers Service (LCS) between the period from March 1988 to June 1992. The evidences from the vast literature regarding female labour force participation and the job-market random matching theory are used to examine the empirical results of the estimators.

The parametric estimator are tighten by the restrictive assumption regarding the link function of the discrete dependent variable and the dummy variables of the discrete covariates. Adding interaction terms improves the performance of the parametric models but encounters other risks like generating multicollinearity problem, increasing the singularity of the data matrix and complicates the computation of the ML function. On the other hand, the mixed data types kernel estimation framework shows an outstanding performance compared with the conventional parametric estimation methods. The kernel functions that are used for the discrete variables, including the dependent variable, in the mixed data types estimation framework, have substantially improved the performance of the kernel estimators. The kernel framework uses very few assumptions about the functional form of the variables in the model, and relay on the right choice of the kernel functions in the estimator.

The outcomes of the kernel conditional density shows that female education level and fertility have high impact on females propensity to work and be in the labour force. The kernel conditional density estimator captures more heterogeneity among the females in the sample than the MNL model due to the restrictive parametric assumptions in the later. The (mixed) proportional hazard framework, on the other hand, missed to capture the effect of the job-market tightness in the job-vacancies hazard rate and produce inconsistent results when the assumptions regarding the distribution of the unobserved heterogeneity are changed. The external kernel hazard estimator overcomes those problems and produce results that consistent with the job market random matching theory. The results in this thesis are useful for nonparametric estimation research in econometrics and in labour economics research.

Key Words: Nonparametric methods, kernel, mixed data types, kernel conditional density, multinomial logit, female labour force participation decision, discrete time hazard, external kernel hazard, sub-survival function, random matching theory.

JEL Classification: C14, C15, C21, C24, C25, C41, C51, C63, J10, J20, J24, J41, J64.

Declaration

Candidate Name: Obbey Ahmed Elamin

Faculty: Humanities

Thesis Title: Nonparametric Kernel Estimation Methods for Discrete Conditional Functions in Econometrics

Declaration:

I declare that no portion of this work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Signed: Date: November 8, 2013

Copyright

Dedication

In the name of Allah, the Most Gracious , the most Merciful

If ye would count up the favours of Allah, never would ye be able to number them; for Allah is Oft-Forgiving, Most Merciful.

chapter An-Nahl verse 18

I am grateful and thankful the most to Allah for giving me the knowledge, the power and the patience to complete this work, and to his prophet of mercy to the mankind, the Prophet Mohammad (pbuh), who lightened and guided me to the right way.

I dedicate this work to my parents:

My father Ahemd Elamin and my mother Hayat Ibrahim.

to my brothers and sisters:

Moaz, Yassir, Osama, Moawia, Omiema, Mai ..... to Amira and to all sisters and brothers in law....

to all my nephews and nieces ... with special dedication to Marwa and Zaid.

Acknowledgements

I dedicated this work to my family and I am still finding it not enough, no words can express how much I am proud of all of them.

I would like to thank my supervisors Len Gill and Martyn Andrews for their continuous encouragement and wonderful help. They have been infinitely patient and supportive all through my PhD period, and incredibly understandable and motivative particularly during the difficult times in this research. I would like also to thank the members of staff at the Economics- University of Manchester, who contributed to this work by suggestions and comments, my special thanks goes to Alastair Hall, Chris Orme, Denise Osborn, Simon Peters, Ken Clark, Ralf Becker and all staff and PhD students in the Econometrics and Applied Economics research group. I would like to thank my colleagues Wasel bin Shadat, Yuyi Li, Mika Kortelainen and Robert O'Neill for the useful suggestions that improved the work in this thesis.

I am grateful to Abdilahi Ali, William Tayler, Laurence Roope, Hormoz Ramian and all friends and colleagues at the Economics, who made my study and time at the Economics nice. I am grateful also to the IT services teams at The University of Manchester, to Simon Hood and Michael Croucher and all other members in the management team of Man2 and RedQueen high performance computer clusters and CONDOR system. Without the supercomputers none of the nonparametric models in this research could be possible. My thanks goes to the Marie Waite and Ann Cronley at the SOSS postgraduate office and to Ann Drozdziak.

I am very thankful to Prof. Jeffrey Racine for his help with R software codes of the conditional density equality test.

I am thankful to my friends and relatives in Sudan who supported me during the past years, and to those who continue to support and encourage me after I started my studies in the UK.

Chapter 1

Introduction and Literature Review

1.1 Parametric Estimation of Discrete Conditional Functions in Econometrics

Estimation of **discrete dependent variable** models is popular in many fields in economics, such as labour economics, welfare economics, transportation, consumer behaviour. The discrete dependent variable model could be a model to estimate the conditional probabilities for multinomial choices, alternatives, to examine ordered outcomes, opinions or ranked choices or to model events in discrete time intervals. There are familiar estimation methods for discrete dependent variable models in conventional parametric estimation frameworks in econometrics. Generally, they are estimated using a **link function**. The link function is a function that maps between the independent variables in the model and the function that being estimated for the discrete dependent variable.

The specification of the link function depends on the function that being estimated, the nature of the discrete dependent variable in the model and the assumptions about the independence and the distribution of the errors in the model. The regular choice of the link function in the multinomial discrete choice models for example is either logit, probit or complementary log-log. The complementary log- log link function is used for modelling in discrete time duration models. There are other types of link functions that are available in the literature in discrete time hazard models estimation. Accordingly, the estimation of discrete dependent variable models is restrictive due to the assumptions that imposed to introduce the link function. The estimated results are often unsatisfactory for econometricians, in term of their consistency with the economic theory. The results are often viewed and interpreted in a very restrictive context that fails to illustrate many aspects in the model.

On the other hand, for the models that include **discrete independent variables** or **categorical independent variables** different treatment is needed. The discrete independent variables are included after converting them to dummy variables and define a base-group. This is considered as a restrictive treatment that limits the ability of the parametric model to capture the effects of those variables appropriately. The dummy variables may explain small amount of heterogeneity in the model. This makes the interpretation of the results of the model valid under very restrictive conditions.

So, including discrete variables in parametric estimation method is problematic due to the illustrated difficulties regard defining them in the model. The restrictions effect is higher when the model include discrete dependent and discrete independent variables, where both link function and dummy variables are needed in the estimator. The parametric estimation of the models is exposed to risks of functional form misspecification, which could result from either an inappropriate specification of the functional form of the variables in the model or from a wrong specification of the link function itself. The methods to check and test the specification of the functional form of discrete dependent variable models are sometimes complicated.

Alternative approaches to estimate discrete dependent variable models, to the link function approach, are very attractive for econometricians. Nonparametric approach is attracting considerable interest nowadays as an estimation framework that does not impose certain specification of the functional form for the objects that being estimated. The functions of the discrete dependent variables are denoted as **discrete conditional functions (DCFs)** in the nonparametric estimation framework. Nonparametric estimation method, particularly the kernel method, provide estimators that can be used to estimate DCFs and handle the discrete independent variables in the model differently. The application of nonparametric estimation methods to the models that include discrete variables, whether as dependent variables, independent variables or both, is slightly attempted in empirical research.

This thesis uses the nonparametric estimation techniques to estimate functions of discrete dependent variables, discrete conditional functions, and compare the results with the conventional parametric discrete dependent variable estimation methods. In this thesis we focus on the multinomial dependent variable case, of discrete unordered choices, and the discrete time to event case, in two empirical examples. The examples that we consider are some microeconometrics applications of discrete dependent variable models.

1.2 Nonparametric Estimation

Nonparametric estimation methods are defined in Racine (2008) as "**statistical techniques that do not require a researcher to specify functional forms for the objects being estimated**". The objective of the nonparametric estimation techniques is to estimate models with as fewest functional form and distribution assumptions as possible. The properties concern the randomness and the smoothness of the data. The early developments of the nonparametric estimation techniques goes back to the 1950s, but in the last decade interest in these techniques has grown sharply. Nonparametric estimation applies a lot more numerical calculations to estimate models than the conventional parametric estimation methods. This needs the researchers to use sophisticated computer routines and wait longer time to receive the results. Due to this hard computation cost that involved in the estimation process, the early nonparametric estimators were little attractive in empirical research.

Interest in nonparametric kernel method has been growing considerably during the last years. Recent books that focus on kernel methods and nonparametric methods in econometrics and statistics include: Ullah (1989) Barnett et al. (1991), Hardie and Linton (1994), Lee (1996), Horowitz (1998), Pagan and Ullah (1999), Härdle et al. (2004), Li and Racine (2007), Li et al. (2009) and Ahamada et al. (2011). Books that present nonparametric estimation techniques in one or more chapters include Powell and Program (1993), Mittelhammer et al. (2000), Ullah et al. (2002), Cameron and Trivedi (2005), Johnston et al. (2007) and Baltagi and Baltagi (2008). Literature survey papers for nonparametric methods in econometrics include Robinson (1988), Ullah (1988a,b), Hardle et al. (1993); Härdle and Manski (1993) and Cai et al. (2009). In addition to the above cited work, there is a vast literature in the empirical applications of the nonparametric estimation techniques in microeconometrics, time series and other fields.

There is early research that compare the performance of nonparametric estimators with their parametric counterparts. Hardle and Marron (1985) compare between the fits of parametric regression with nonparametric regression and show that there are visible differences between the fits that generated by each technique. Nicol (1993) compares empirically between parametric and nonparametric regressions in estimating an Engle Curve, and finds that nonparametric regression is more flexible than parametric regression. Härdle et al. (2004), Li and Racine (2007), Racine (2008) and Cai et al. (2009) provided more examples for similar comparisons between parametric and nonparametric techniques and agree in the advantages of the nonparametric estimation. Generally nonparametric approach is recommended when the data is not well behaved, for example when there is a fat tail behaviour or sharp skewness occurring in the data, or in the models for rare events. Nonparametric estimation methods are preferred also when the parametric estimation fails to give satisfactory results that match with the economic theory, as in the example of the conditional hazard of job vacancies in Chapter 4 in this thesis, and in models with unobserved heterogeneity. The nonparametric estimation approach in empirical research is attractive when the model include only a few variables, the sample size is sufficiently large, the functional form of the parametric models is complicated and problems like endogeneity or random effects are not anticipated in the model.

The recent development in the nonparametric estimation techniques is encouraged by the remarkable development in computer technologies and software engineering. Which contributed to make the work with those techniques easier and encouraged the research in this field. Currently, there are large number of new nonparametric estimators that are available, or suggested in the literature, in econometrics. The volume of the literature concerning those techniques is growing exponentially for years. The interest to use nonparametric estimation techniques empirically in econometrics, following the revolution in the theoretical development, is now high. In this thesis we are motivated by the development in kernel method estimators, and our aim is to apply some of the new estimators in empirical research in applied microeconometrics. We focus on kernel method with fixed bandwidth, that is defined in the next subsection, for its advantages on smoothing discrete variables efficiently. Our interest is to compare the empirical results of the kernel method with the conventional parametric discrete dependent variable methods.

1.2.1 The Kernel Method with Fixed Bandwidth

Kernel method with fixed bandwidth (KMWFB) is a nonparametric estimation technique that has many advantages over other nonparametric estimation methods in econometrics. The method is popular because it is easy to handle theoretically and empirically, allows to smooth discrete (categorical) variables and construct estimators for mixed types of variables and allows to construct smoothing tests, whether for testing the significance of the variables in the model or to test the specification of the model. Kernel method aims to estimate densities and/or distribution functions directly by smoothing the variables. The kernel is a smoothing function that weights the observations in the sample using specific rule that depends on the distance between the observations. To illustrate the KMWFB and its approach to estimate the density function we start by the simplest kernel function that smooths sample observations, the **naïve kernel** of Van Ryzin (1973) for continuous variables. The density estimation using the naïve kernel is known as the **histogram method**. The naïve kernel is just a simple indicator function that applies a simple rule to the variables in the model.

Let X_i^c, $1, 2, ..., n$, be an iid continuous random variable that drawn from unknown distribution $f(x^c)$. The superscript c denotes that X is a continuous variable. Our objective is to estimate the density of X^c from a sample of n observations. The fundamental information about the density of X is that it is a smooth function that integrate to 1. The density, as a continuous curve has a property that for any nearby values x_1^c and x_2^c in the domain of f, the density $f(x_1^c)$ has near value to $f(x_2^c)$. As $x_1^c \rightarrow x_2^c$ the density exhibit that $f(x_1^c) \rightarrow f(x_2^c)$. The histogram method utilise this characteristic in the estimation procedure. The sample observations that are close in values in fixed ranges known as the **binwidth,** b, are taken to estimate the density. The observations are set into bins in an attempt to estimate the density. The estimator takes the following form, see Härdle et al. (2004, page 22),

$$\widehat{f}_b(x^c) = \frac{1}{nb}\sum_{i=1}^{n}\sum_{j}^{k} I(X_i^c \in B_j) I(x^c \in B_j), \tag{1.1}$$

where $I(X_i^c \in B_j)$ is an indicator function that equals 1 if X_i^c is included in the j^{th} bin, b is the binwidth and k is the number of bins. The naïve kernel, $I(X_i^c \in B_j)$, equally weights the observations in each bin.

The disadvantage of the histogram method is that it is not smooth and fully dependent on a choice of an origin point, x_o^c, that needs to be determined in advance, and on the value of the binwidth. Different choices of these two parameters lead to a substantial changes in the density that being estimated. Arbitrary choice of b and x_o^c is against the objectives of nonparametric estimation, because they are the key elements in the model and the arbitrary choice ignores to use the data in the sample. In other word, arbitrary choosing the value of b is not a data driven estimate. This contradicts with the rule in the nonparametric estimation that says "let the data speak for themselves", Härdle et al. (2004).

Despite that there are data driven methods to estimate the optimum value of b in the literature, like the technique that based on minimizing the **mean integrated square error (MISE)**, the histogram method is still considered as a rough technique that produces inaccurate results. Ullah (1988b) mentions that the histogram method is discontinuous and complicated in multivariate density estimation, Wegman (1972b) shows inaccuracy of the histogram method using simulation methods and concludes that the histogram density estimators have error in the rate of convergence. The same results are supported in Kumar and Markman (1975). Additionally, the histogram method needs a large sample size and it is badly affected by the outliers.

Kernel Univariate Density Estimation

The kernel method is suggested by Rosenblatt (1956) to estimate univariate probability density functions for continuous random variables. The kernel method overcomes many of the disadvantages of the histogram method. The objective to estimate a strong smooth density function using the observations in the sample is achieved in the kernel estimator. The kernel is defined as a weighting function that weights the observations in the sample, X_i^c, based on their distance from a specific value x^c. Then x^c is known as the **point of smoothing** or the **point of estimation**. X_i^c are weighted relative to x^c in a fixed range known as the **window width** or the **bandwidth**. The weights that are given by the kernel function to the observations in the sample are known as the **local weights**. The basic density estimator is the sample average of the local weights that given by the kernel for all the observations in the sample. For a continuous univariate density function, $f(x^c)$, the kernel with fixed bandwidth estimator at the value x^c using sample observation $\{X_i^c\}_{i=1}^n$ is given as follows, see Pagan and Ullah (1999, page 9):

$$\widehat{f}(x^c) = \frac{1}{n\widehat{h}}\sum_{i=1}^{n} k\left(\frac{X_i^c - x^c}{\widehat{h}}\right) = \frac{1}{n\widehat{h}}\sum_{i=1}^{n} k\left(\widehat{z}_i\right), \tag{1.2}$$

which known as the Rosenblatt-Parzen density estimator. $h \geq 0$ is the fixed bandwidth of the kernel function, equivalent to the binwidth b in the histogram method.

The bandwidth, h, has fixed value that does not change irrespective to whether the point of estimation, x^c, is at the centre or at the boundary of the distribution. If x^c is at a boundary the number of the sample observations in that range is likely to be small. Then keeping the window width fixed at all values of x^c makes the estimation of the density at the boundaries biased and inconsistent, since this ignores the data density at those points, see Gasser et al. (1985) and Karunamuni and Alberts (2005). To overcome this problem researches suggest kernel method with variable bandwidth, also known as the k nearest neighbour (k-nn) method, that use bandwidth that vary with x^c, denoted as h_{x^c}. This reduces the bias in the kernel estimator. However, the advantage of KMWFB are sill highly attractive, where it allows to smooth variables of different type jointly and construct smoothing significance tests.

The kernel function weights the observations in an interval of $\pm h$ around x^c. The simplest form of weights is given by the Uniform (the naïve) kernel, which is simply a function that gives equal weights $\frac{1}{2}$ for all the observations inside the interval $[x^c - 1, x^c + 1)$, and zero weights for all the observations outside this interval, Härdle et al. (2004). Other kernel functions apply different types of weights. Each kernel function, however, applies a unique type of weights, some of which have highly sophisticated formulas. The general rule in all kernel functions is that; the closer the observation in the sample to x^c, the higher the weight that is given to that observation by the kernel function. So, in the density estimator the observations that are near x^c and inside the interval $[x - h, x + h)$ have higher weights than the far observations inside the interval, and all the observations outside the interval are given zero weights. However, some kernel functions weight the observations outside the interval as well, for example the Gaussian Kernel function, where the observations outside the interval are not given zero weights. Table 1.1 presents some kernel functions for continuous variables, the indicator $I(|z| < 1)$ means that the observations outside the interval $[x - h, x + h)$ are given 0 weight.

The kernel function is a symmetric function and satisfies the following consistency conditions;

1. $\int k\,(z)\,dz = 1.$

2. $\int z^s k\,(z)\,dz = 0.$

3. $\int z^r k\,(z)\,dz = \tau_r \neq 0.$

When $r = 2$ the kernel is known as a first order kernel. The second condition ensures the symmetry property. When $r > 2$ the kernel becomes known as a higher order kernel but the symmetry condition still holds.

Standardised values of the observations are used in the kernel function. In this context, the bandwidth plays the role of the dispersion measure, the standard divination. The kernel function weights the standardised value $z_i = \frac{X_i^c - x^c}{h}$ for each observation in the sample when the density is estimated at x^c. For each observation in the sample the area under the kernel function is $\frac{1}{n}$, so that the total area under all the observations is 1.

The derivatives of the Gaussian kernel function are easy to find, this facilitates the estimation of the derivatives and the partial derivatives of the density estimator. High order kernels reduce the bias in the kernel density function estimator but require more numerical calculations, Jones and Signorini (1997). This fact makes higher order kernels functions preferred for density estimation in small samples. Discussion about the effect of the order of the kernel functions on the precision of the kernel density estimators is found in Marron (1994), Jones (1995), Jones and Signorini (1997), Gasser et al. (1985), Ahmad and Ran (2004) and Greco et al. (2009). More advanced discussion with infinite order kernel functions and flatted-top kernel functions is presented in Politis and Romano (1999). Recent developments in nonparametric estimation adds additional advantages for the higher order kernel functions, where Li and Racine (2007) show that high order kernel functions allow for constructing the smoothing significance and specification tests for nonparametric estimators better than low order kernel functions. High order kernel functions ensure the asymptotic properties of the smoothing test and increase the convergence rate, but makes the test difficult empirically due to the tough computation of the test statistic and its empirical bootstrap distribution. More discussion about kernel specification tests is given in the discussion of discrete choice models specification tests in Chapter 2.

Table 1.1: Kernel functions for continuous variables[1]

Kernel Function for Continuous Variables[2]	
Uniform (naïve) Kernel	$\frac{1}{2} \times I(\lvert z\rvert < 1)$
Triangle	$(1 - \lvert z\rvert)\, I(\lvert z\rvert < 1)$
Epanechnikov (or Quadratic)	$\frac{3}{4}\left(1 - z^2\right) \times I(\lvert z\rvert < 1)$
Gaussian (or normal)	$(2\pi)^{-\frac{1}{2}} \exp(-z^2/2)$
Quadratic	$\frac{15}{32}\left(3 - z^2\right)^2 \times I(\lvert z\rvert < 1)$
Triweight	$\frac{35}{32}\left(1 - z^2\right)^3 \times I(\lvert z\rvert < 1)$
Fourth order Gaussian	$\frac{1}{2}(3 - z)^2 (2\pi)^{-\frac{1}{2}} \exp(-z^2/2)$
Fourth order Quadratic	$\frac{15}{32}\left(3 - z^2 + 7z^4\right)^3 \times I(\lvert z\rvert < 1)$

[1] Source Cameron and Trivedi (2005), Härdle (2004).
[2] $z = \frac{X^c - x^c}{h}$.

The kernel method is a viable technique to estimate density functions, works better for symmetric distributions and is extendible to estimate many type of functions, like conditional density, conditional hazard, regression and distribution functions. The kernel method is used also to estimate densities of complex variables like the truncated and censored variables. During the 1970's there was intensive research in kernel methods. Simulation methods are used to examine the properties of the kernel estimators and compare them with other nonparametric methods of density estimation. Or to compare different methods of bandwidths estimation for kernel density estimator. For example the work of Wegman (1972b,a) and Kumar and Markman (1975). Fryer (1977) compares the asymptotic properties of a number of nonparametric methods of density function estimation including the kernel method and finds that kernel density estimators have 'infinite support' and good rate of convergence compared with the histogram method estimator.

Fryer (1977) additionally illustrates that the histogram density estimator has "error rate of convergence", which is a term that used to describe the convergence that is higher than ($O(n^{-0.8})$), see Fryer (1977, Table 2, page 23). In a comparison study between the kernel and the series methods, Anderson (1969) applies the Monte Carlo simulation technique to illustrate that kernel method produces better estimated densities in term of the mean integrated square error (MISE) measure. Scott and Factor (1981) apply Monte Carlo simulation also to compare the kernel density estimator with the k-nn methods density estimators. The results show dissimilar outcomes for the kernel method and the series method in the large sample. Scott and Factor (1981) recommend the kernel method, and argue that kernel method is easier to be computed and workable more than the series and k-nn methods.

The convergence rate of the univariate continuous density under the consistency conditions above is: as $h \to 0$, $nh \to \infty$ and $n \to \infty$

$$\widehat{f}(x^c) - f(x^c) = O_p\left(n^{-\frac{2}{5}}\right). \tag{1.3}$$

Among the nonparametric methods of density estimation the histogram method has the slower rate of convergence. The rate of convergence of a correctly specified parametric density estimator is, Racine (2008, page 12):

$$\widehat{f}(x) - f(x) = O_p\left(n^{-\frac{1}{2}}\right). \tag{1.4}$$

So the nonparametric technique has a slower rate of convergence than that for the correctly specified parametric estimator, $n^{-\frac{1}{2}}$.

The asymptotic normality result for the kernel density estimator is given by:

$$\sqrt{nh}\left(\widehat{f}(x^c) - f(x^c)\right) \xrightarrow{d} N\left(0, f(x^c). \int k^2(v)\,.dv\right) \quad \text{as } n \to \infty. \tag{1.5}$$

The 95% confidence intervals for the density function are then estimated as follows:

$$\widehat{f}(x^c) \pm 1.96\sqrt{\frac{f(x^c).\int k^2(v).dv}{nh}}. \quad (1.6)$$

Univariate Probability Mass Function Estimation

To estimate probability mass functions for discrete variables the frequency method is used, but considered as a simple method that equivalent to the histogram method in continuous variables density estimation. The frequency method split the data into cells to estimate the mass probability function. Let X_i^d, $1, 2, ..., n$, be an iid discrete random variable that has an unknown probability mass function $p(x^d)$, the superscript d denotes that x is a discrete variable. The variable x^d is assumed to have finite and limited support in $\{0, 1, 2, ..., c-1\}$. The probability mass function estimate is given by the follows formula, Li and Racine (2007, page 116):

$$\widehat{p}(x^d) = \frac{1}{n}\sum_{i=1}^{n} I(X_i^d = x^d), \quad (1.7)$$

which is the relative frequency at x^d.

The frequency method produces unbiased and asymptotically consistent estimators if the mass probability function has few mass points, but needs large sample to operate well. If the number of values in the discrete variable is large, the accuracy of the frequency method deteriorates dramatically. If the number of cells, on the other hand, exceeds the sample size, the frequency method collapses completely. Additionally, the frequency method is inapplicable for the multivariate probability functions.

The extension of KMWFB to estimate probability mass functions of categorical variables is provided in Aitchison and Aitken (1976). They propose an estimator in the univariate case that takes the following form, Racine (2008, page 17)

$$\widehat{p}\left(x^d\right) = \frac{1}{n}\sum_{i=1}^{n} l\left(X_i^d, x^d, \widehat{\gamma}\right). \quad (1.8)$$

Where $l\left(X_i^d, x^d, \gamma\right)$ is a weighting function that depends on γ (the window width or the bandwidth). The smoothing parameter, γ, take a value in $[0, 1]$ range and depends on the number of values in the support of X^d variable, c. The consistency conditions for the discrete variable kernel functions are similar to that for the continuous variable kernels, where the discrete probability function is estimated under the following conditions:

For any discrete variable that takes values in the range $\{0, 1, ..., c-1\}$, *the kernel (weighting) function for this variable is defined as, Wang and Van Ryzin (1981);*

$$\sum_{x^d=0}^{c-1} l\left(X_i^d, x^d, \gamma\right) = 1. \quad (1.9)$$

Where $l\left(X_i^d, x^d, \gamma\right) \geq 0$, *for every* $x^d = 0, 1, ..., c-1$, $i = 1, 2, ..., n$ *and* $\gamma \in [0, 1]$.

To ensure the asymptotic properties, the discrete kernel function should satisfy the following conditions, Wang and Van Ryzin (1981), Ahmad and Cerrito (1994):

1. $\gamma \to 0$ as $n \to \infty$.

2. $\gamma/\sqrt{n} \to 0$ as $n \to \infty$, and $l\left(X_i^d, x^d, \gamma\right)$ has a continuous first derivative at 0.

The first condition ensures that $\widehat{p}\left(x^d\right)$ converge in probability to $p\left(x^d\right)$. While the second condition ensures that the estimator has an asymptotic normal distribution:

$$\sqrt{n}\left\{\widehat{p}\left(x^d\right)\left(1-\widehat{p}\left(x^d\right)\right)\right\} \xrightarrow{d} N\left(0, p\left(x^d\right)\left(1-p\left(x^d\right)\right)\right). \tag{1.10}$$

The estimator has the same structure for both unordered variable and ordered variable probability mass density functions, but the type of the discrete kernel function, $l(\cdot,\cdot,\cdot)$, is different. There are large number of discrete kernel functions that have been developed in the literature in the work of Habbema et al. (1978), Titterington (1980), Wang and Van Ryzin (1981), Aitken (1983) and Titterington and Bowman (1985). Tables 1.2 and 1.3 present examples of the kernel functions for unordered and ordered variables respectively. The kernel functions in the tables are collected from different sources as shown in the footnotes to the table.

The discrete kernel function weights the categorical variable by matching each value in the sample, X_i^d, with the value at the point of smoothing, x^d. The difference between the unordered kernel function and the ordered kernel function is on the method of weighting the pairs X_i^d and x^d. The unordered kernel function gives only two weights, the first weight is for the matched pairs X_i^d and x^d, which is equivalent to the case when $X_i^d - x^d = 0$. The second weight is for the dismatched pairs X_i^d and x^d, i.e. when $X_i^d - x^d \neq 0$. The ordered kernel function, in contrast to the unordered kernel, gives weights that depend on the distance $\left|X_i^d - x^d\right|$. Which means that the ordered kernel function give c different weights to the observations in the sample. The absolute values of the distances are taken because the kernel function is symmetric.

Table 1.2: Kernel function for discrete unordered variables[1]

Aitchison and Aitken (1979) [1]	$l\left(X_i^d, x^d, \gamma\right) = \begin{cases} 1-\gamma & \text{if } X_i^d = x^d \\ \gamma/(c-1) & \text{if } X_i^d \neq x^d \end{cases}$
Aitken kernel (1976) [2]	$l\left(X_i^d, x^d, \gamma\right) = \begin{cases} 1 & \text{if } X_i^d = x^d \\ \gamma & \text{otherwise} \end{cases}$

[1]Li and Racine (2007).[2]Li et al. (2008).

Table 1.3: Kernel function for discrete ordered variables

Aitchison and Aitken[1]	$l\left(X_i^d, x^d, \gamma\right) = \begin{cases} \binom{c}{j} \gamma^j (1-\gamma)^{c-j} & \text{if } \lvert X_i^d - x^d \rvert = j \\ 0 & \text{if } \lvert X_i^d - x^d \rvert \neq j \end{cases}$	
Modified Aitchisonand Aitken[1]	$l\left(X_i^d, x^d, \gamma\right) \propto \begin{cases} 1 & \text{if } X_i^d = x^d \\ \gamma^{\lvert x^d - x^d \rvert} & \text{if } X_i^d \neq x^d \end{cases}$	
Habbema kernel[4]	$l\left(X_i^d, x^d, \gamma\right) \propto \gamma^{\lvert X_i^d - x^d \rvert^2}$	$\gamma \in [0,1]$
Modified Habbema kernel[2]	$l\left(X_i^d, x^d, \gamma\right) \propto \gamma^{\lvert X_i^d - v^d \rvert^\lambda}$	$\gamma \in [0,1]\,, \lambda \in [0,\infty]$
Aitken kernel[3]	$l\left(X_i^d, x^d, \gamma\right) = \begin{cases} \gamma & \text{if } X_i^d = x^d \\ (1-\gamma) \big/ 2^{\lvert X_i^d - x^d \rvert} & \text{if } X_i^d \neq x^d \end{cases}$	$\gamma \in [0,1]$
Geometrical [3]	$l\left(X_i^d, x^d, \gamma\right) = \begin{cases} 1-\gamma & \text{if } X_i^d = x^d \\ \frac{1}{2}(1-\gamma)\gamma^{\lvert X_i^d - x^d \rvert} & \text{if } X_i^d \neq x^d \end{cases}$	$\gamma \in [0,1]$
Titterington (1980)[4]	$l\left(X_i^d, x^d, \gamma\right) = \begin{cases} 1-\gamma & \text{if } X_i^d = x^d \\ (1-\gamma)\, g_{ji} & \text{if } X_i^d \neq x^d \end{cases}$ $\sum_{j=0}^{c-1} g_{ij} = 1$ for all i, and $g_{ij} \geq 0$	$\gamma \in [0,1]$
Uniform [5]	$l\left(X_i^d, x^d, \gamma\right) = \begin{cases} \frac{1}{2}(1-\gamma)/c-1 \quad \text{if } \lvert X_i^d - x^d \rvert = j \\ 1-\gamma \quad \text{if } X_i^d = x^d \\ 0 \text{ otherwise} \end{cases}$ $\gamma \in [0,1]$ and $j \leq c-1$	

[1]Li and Racine (2007), [2]Tutz and Pritscher (1996), [3]Li et al. (2008), [4]Titterington and Bowman (1985), [5]Wang and Van Ryzin (1981).
Modified Aitchisonand Aitken, Habbema and Modified Habbema kernel functions need to be normalised to satisfy the condition $\sum_{x^d=0}^{c-1} l\left(X_i^d, x^d, \gamma\right) = 1$.

Many studies compare the accuracy and the asymptotic behaviour of the kernel functions, but with minor distinction between unordered and ordered kernel functions. Examples of such a comparison is found in the work of Titterington (1980). Comparison between ordered kernel functions is available in Titterington and Bowman (1985), which illustrates that Habbema kernel function, in Table 1.3, is more efficient compared with the Uniform and Titterington kernel functions. Li and Racine (2003) modified the Aitchison and Aitken (1976) ordered kernel to the Modified Aitchison and Aitken kernel to use it in the mixed data smoothing framework, that is defined in the next sub-section. The original ordered kernel function of Aitchison and Aitken (1976) cannot smooth out (drop out) the irrelevant variables in the kernel density estimator when the cross validation method is used to estimate the bandwidths. This characteristic is illustrated more in Li and Racine (2007). The modified version of the kernel function in Table 1.3 seems like a special form of Modified Habbema kernel, only with $\lambda = 1$. Li and Racine (2007) claim that the modification of the kernel is necessary only for the independent variables in the model, if the discrete variable is a dependent variable it is not necessary to use a discrete kernel function that capable to drop out irrelevant variables to smooth it as Li and Racine (2007) claim.

Nonparametric kernel estimators with discrete variables have a better convergence rate. The rate of convergence of a density estimator of a discrete variable equals the rate of convergence of a correctly specified parametric estimator, where

$$\widehat{p}\left(x^d\right) - p(x^d) = O_p\left(n^{-\frac{1}{2}}\right), \tag{1.11}$$

1.2.2 Multivariate Kernel Density Estimation with Mixed Types Variables

The multivariate kernel density is estimated by using the product of the univariate kernel functions. Therefore, estimating the multivariate density of a same type variables is easier than estimating a multivariate density of mixed types variables. The estimator of a multivariate density function of q continuous variables takes the following form, rearranged from Li and Racine (2007, page 25):

$$\widehat{f}\left(\mathbf{x}^c\right) = \widehat{f}\left(x_1^c, x_2^c, ..., x_q^c\right) = \frac{1}{n\widehat{h}_1...\widehat{h}_p}\sum_{i=1}^{n}\prod_{s=1}^{q}k\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right), \tag{1.12}$$

For a multivariate density of p discrete variables the formula is the following, Li and Racine (2007, page 127):

$$\widehat{p}\left(x^d\right) = \widehat{p}\left(x_1^d, x_2^d, ..., x_p^d\right) = \frac{1}{n}\sum_{i=1}^{n}\prod_{r=1}^{p}l\left(X_{ir}^d, x_{ir}^d, \widehat{\gamma}_r\right). \tag{1.13}$$

The estimation framework of objects that include mixture of continuous and discrete variables using KMWFB is known as "**mixed data types (MDT) kernel estimation framework**". The framework is developed by Jeffrey Racine, Qi Li and Peter Hall and other researchers, in a series of papers that propose estimators for multivariate density function, multivariate conditional density, nonparametric regression with mixed types regressors and many other estimators. MDT kernel estimation framework opens the door for wider applications of the nonparametric estimation techniques in econometrics. Most of the objects that researches aim to estimate are functions of mixed discrete and continuous variables. The framework contributes directly on the development of kernel estimators for discrete conditional functions, the functions of discrete dependent variable. Additionally, it allows to have a nonparametric kernel estimators that be the nonparametric counterpart for the discrete choice models like the multinomial logit and multinomial probit, the ordered choice models or discrete time hazard rate. The MDT framework is developed under KMWFB, i.e the kernel functions for continuous variables must use the fixed bandwidth rule to smooth the mixed types variables correctly. Then KMWFB is the broad nonparametric estimation method that we use. In this literature review we will present some of the estimators that are suggested in this framework.

The multivariate density function is the easiest MDT estimator. The multivariate density might include variables of one type or a mixed of continuous and discrete variables types. Early attempt of estimating a multivariate density function for mixed discrete and continuous variables is created by Ahmad and Cerrito (1994), where they use a uniform and a geometric discrete kernel functions to estimate a bivariate distribution of one continuous and one discrete variables. For more than one discrete variable the choice of the kernel function in the multivariate kernel density estimator was problematic, until the development of the MDT smoothing technique. Li and Racine (2003) show that the frequency method is inappropriate in the mixed type variables density estimator, as it splits the data into large number of cells, which may exceed the sample size if the model includes many variables. In the papers Li and Racine (2003), Racine et al. (2004), Li and Zhou (2005), Racine et al. (2006), which are part of the series of papers that demonstrate the smoothing method for MDT. Which is illustrated more also in Li and Racine (2007), the authors suggest the **Aitchison and Aitken** kernel function for unordered variables, and the **Modified Aitchison and Aitken** kernel function or the **Geometrical Kernel** function for ordered variables.

For a multivariate density that includes p continuous variables and q discrete variables, let the sample include $\{\mathbf{X}_i^c, \mathbf{X}_i^d\}_{i=1}^n$, where $\mathbf{X}_i^c$ is the vector of the sample observations of the continuous variables for individual i, $\mathbf{X}_i^c = [X_{1i}^c, \cdots, X_{qi}^c]$, and $\mathbf{X}_i^d$ is the vector of the discrete variables, $\mathbf{X}_i^d = [X_{1i}^d, \cdots, X_{pi}^d]$. To estimate the mixed density at the vector of values $\mathbf{x} = [\mathbf{x}^c, \mathbf{x}^d]$ the estimator takes the following form, rearranged from Li and Racine (2007).

$$\begin{aligned}
\widehat{f}(\mathbf{x}) &= \widehat{f}\left(x_1^c, ..., x_q^c, x_1^d, ..., x_p^d\right) \\
&= \frac{1}{n\widehat{h}_1...\widehat{h}_q} \sum_{i=1}^{n} \prod_{s=1}^{q} k\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right) \prod_{r=1}^{p} l\left(X_{ir}^d, x_{ir}^d, \widehat{\gamma}_r\right) && (1.14) \\
&= \frac{1}{n\widehat{h}_1...\widehat{h}_q} \sum_{i=1}^{n} \mathbf{W}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}}), && (1.15)
\end{aligned}$$

where $\mathbf{W}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}}) = \prod_{s=1}^{q} k\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right) \prod_{r=1}^{p} l\left(X_{ir}^d, x_{ir}^d, \widehat{\gamma}_r\right)$ is the product of the univariate kernel functions of the mixed types variables in the model at observation i in the sample, and $\widehat{\mathbf{h}}$ is the vector of the estimated bandwidths of the variables in the model, $\widehat{\mathbf{h}} = [\widehat{h}_1, \cdots, \widehat{h}_q, \widehat{\gamma}_1, \cdots, \widehat{\gamma}_q]$.

In the multivariate case the rate of convergence is $O_p\left(n^{-\frac{1}{q+4}}\right)$ for the multivariate density with continuous variables only. In the models with discrete and continuous variables the continuous variables control the convergence rate.

1.2.3 Estimation of the Bandwidths

Kernel density estimators depend crucially on the smoothing parameters (the $h's$ in the continuous kernel and the γ's in the discrete kernels). The precision of the nonparametric kernel density estimator deteriorates dramatically with the inappropriate choices of the smoothing parameters. In the kernel nonparametric estimation method the choice of the kernel functions in the estimator, by construction, is not as sensitive as the choice of the bandwidths. The bandwidth values influence on the precision of the kernel estimates more, i.e. on the estimated standard error of the density and on the convergence rate to the true density. Most of the calculation cost in kernel density estimation is paid in the estimation of the bandwidths. Data-driven methods to estimate the bandwidths require hard calculations. The numerical calculations become harder when the sample size increases, the number of variable in the model increase or higher order kernel functions are used. However, higher order kernel functions speed up the convergence rate to the true density as shown by Li and Racine (2007).

There are two main approaches: the bandwidth approximation method which approximates the true theoretical bandwidths of the kernel function, and data driven bandwidth estimation methods. Data driven methods estimate the bandwidths by optimising an objective function that trades off between the variance and the bias of the kernel density estimator. The variance and the bias of the continuous univariate kernel density estimator in Eq (1.2) is shown by Racine (2008, page 10) as follows:

$$Bias\ \widehat{f}(x^c) \approx \frac{h^2}{2} f^{(2)}(x^c) \int z^2 k\left(z\right) dz \quad , \qquad var \widehat{f}(x^c) \approx \frac{f(x^c)}{nh} \int k^2(z) dz, \tag{1.16}$$

where $f^{(2)}(x)$ is the second derivative of the density function.

When the consistency conditions that $h \to 0$, $nh \to \infty$ as $n \to \infty$ are satisfied, the bias in the kernel density estimator is eliminated. The optimum bandwidth is the value that minimises the **integrated mean square error (IMSE)**, Li and Racine (2007, page 13):

$$IMSE(\widehat{f}) \stackrel{\text{def}}{=} \int E\left[\widehat{f}(x^c) - f(x^c)\right]^2 dx, \tag{1.17}$$

which gives

$$h_{opt} = \left\{ \frac{\int k^2\left(z\right) dz}{\left[\int z^2 k\left(z\right) dz\right]^2 \left[\int f^{(2)}(x) dx\right]^2} \right\}^{-\frac{1}{5}} .n^{-\frac{1}{5}} = c_0 n^{-\frac{1}{5}}. \tag{1.18}$$

Then, the optimum bandwidth is a function on the second derivative of the true density, which is unknown in the model. Bandwidth approximation methods use underlying assumptions about the true density. This considered inconsistent with the objectives of nonparametric estimation, but they may be attractive in a model with a large number of variables or large sample size. Generally, the bandwidth estimation method can be one of the following methods.

Trial and Error Approach (Graphical Selection Approach)

This method is shown in Pagan and Ullah (1999), and incorporated in many software packages as an easy and arbitrary method that uses graphical presentation. The Trial and Error approach is fully dependent on an arbitrary choice of h. This could be chosen after studying of a number of plots of $\widehat{f}(x^c)$ with x^c with different values of h. Hence, the Trial and Error method is applicable only when the sample size is small and the model includes few variables, one or two variables only. For multivariate densities the Trial and Error approach becomes very difficult and ineffective.

Plug-in Method

The Plug-in method, introduced by Woodroofe (1970), assumes that the variable in the model follows a certain density function, then uses the formula in Eq (1.18) to obtain an initial pilot value of h.

The disadvantage of the Plug-in method is that it is not fully nonparametric and inconsistent with the objectives of the nonparametric estimation. The assumption regarding the distribution of the variable is not plausible in kernel estimation.

Rule of Thumb Method

Probably the Rule of Thumb is one of the oldest methods to estimate the smoothing parameter. The Rule of Thumb method is introduced and recommended by many researches during the early stages of the development of the nonparametric kernel estimation techniques, for example by Deheuvels (1977) and Silverman (1986). A pilot value obtained by one of the above methods, the Trial and Error method or the Plug-in method is used, as a smoothing parameter itself. So it conflicts with the objective of the nonparametric method, because it does not include any searching process for the optimal bandwidth. Jones et al. (1996) in their brief survey of smoothing parameters selection for density estimation, mentioned that the Rule of Thumb method has the disadvantage of smoothing out some important features of the data.

However, when the cost of the numerical calculation increase, the Rule of Thumb method offers a solution, particularly for the large sample sizes, since the true bandwidths converge in probability to some known values.

Cross-Validation Methods

Cross validation (CV) methods are a set of data-driven bandwidth estimation techniques that attract most of the attention in the later research in nonparametric estimation. They aim to estimate the smoothing parameter of the kernel function automatically from the sample, and defined as “extremum” estimators that are generated by optimizing a loss objective function on the true density. The **integrated square error (ISE)**, mean integrated square error (MISE) which equals the integrated mean square error (IMSE), the **weighted integrated mean square error (WIMSE)** and the **asymptotic integrated mean square error (AIMSE)** are the commonly used functions in this context.

Cross validation method is introduced to the nonparametric literature separately by Duin (1976) and Habbema, Hermans and Van Der Burgt (1974), and discussed intensively in Scott and Factor (1981), Rudemo (1982), Hall (1982), Bowman (1984), Stone (1984), Silverman (1986), Eubank (1988), Härdle (1989) and Scott and Terrell (1987). The asymptotic property of the bandwidths that are estimated by the CV method are examined by Hall and Marron (1987), Park and Marron (1990), Hall et al. (1992), Jones et al. (1996), Chiu (1991, 1992, 1996). The latter compares different methods of bandwidth estimation for density functions. The development of CV methods has two breakthroughs; first, in the work of Sarda (1993) who introduces the so-called **leave-one-out (LOO)** method that improved the quality of the estimated bandwidths substantially, and second in the work of Sain et al. (1994) that extends the use of the cross validation method to multivariate densities.

In the following a review of some CV methods that are used in the kernel density estimation.

Least Squares Cross-Validation This method is proposed by Fryer (1977), Rudemo (1982) and Bowman (1984). The **least squares cross validation (LSCV)** optimum bandwidths are estimated by minimising the integrated squared difference between the true function $f(\mathbf{x})$ and kernel function $\widehat{f}(\mathbf{x})$ as follows:

$$\min_{h_1\dots.h_q,\gamma_1\dots.\gamma_p}\left\{\int\left[\widehat{f}(\mathbf{x})-f(\mathbf{x})\right]^2 d\mathbf{x}^c\right\}. \tag{1.19}$$

In the case of multivariate density function for q continuous variables the LSCV formula takes the following form, compiled from Li and Racine (2007, page 27) and Pagan and Ullah (1999, page 51).

$$\begin{aligned} CV_f(\mathbf{h}) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i,j=1}^{n}\prod_{s=1}^{q}h_s^{-1}k\left(\frac{X_{is}-X_{js}}{h_s}\right) \\ &\quad+\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i,j=1}^{n}\prod_{s=1}^{q}h_s^{-1}\left(k\circ k\right)\left(\frac{X_{is}-X_{js}}{h_s}\right), \end{aligned} \tag{1.20}$$

where $(k \circ k)$ is the convolution of the kernel with itself. For a multivariate function with q continuous variables and p discrete variables the LSCV objective function is presented in (Li and Racine; 2003, page 270)

The LSCV method is recommended for kernel density estimation in Hall (1983), Burman (1985) and Stone (1984), with the asymptotic properties of the bandwidths discussed in Hall et al. (1992). The problem with the LSCV method as stated by Racine (2008) is that it is “sensitive to the presence of rounded or disretized data and to small scale effects in the data". But the bandwidths converge to the optimum bandwidths. Li and Racine (2003) recommend the LSCV method to estimate the bandwidths for the density functions with mixed variables type. It is recommended as first choice if the model does not include many variables or the sample size is not large, due to the tail problem that is associated with the alternative method, the maximum likelihood cross-validation MLCV method, that is illustrated below.

Maximum Likelihood Cross-Validation The bandwidths can be estimated using the **maximum likelihood cross-validation (MLCV)** method, which is constructed from the **kullback-leibler distance** between the kernel density and the true density, it is known also as the **kullback-leibler cross-validation** method. The MLCV method finds the optimum bandwidth values by maximising the following objective function, Racine (2008, page 15):

$$L = \sum_{i=1}^{n} \log \widehat{f}_{-i}(\mathbf{X}_i) = \sum_{i=1}^{n} \log \left\{ \frac{1}{(n-1)\,\mathbf{h}^c} \sum_{j \neq i, j=1}^{n} \mathbf{W}(\mathbf{X}_i, \mathbf{X}_j, \widehat{\mathbf{h}}) \right\}, \tag{1.21}$$

which is the sample analogue of the following loss function, Pagan and Ullah (1999, page 51):

$$I\left(\widehat{f}, f\right) = \int \widehat{f}(\mathbf{x}) \log \left\{ \frac{f(\mathbf{x})}{\widehat{f}(\mathbf{x})} \right\} dx,$$

$\widehat{f}_{-i}(\mathbf{X}_i)$ is known as the **leave-one-out LOO** kernel density estimator. The density in the LOO estimator is estimated for observation i in the sample using the $n-1$ other observations. The LOO procedure is crucial for CV, as it developed from the Jackknife bootstrap method to avoid having zero value of the objective function. Without the LOO procedure the cross validations methods collapse completely. The main disadvantage of the maximum likelihood cross-validation method is that it is badly affected by the fat tailed distributions, like the Cauchy distribution, where MLCV oversmooths the kernel distributions.

Comparisons of bandwidth estimation methods for unconditional density had wide discussion in nonparametric econometrics. Loader (1999) in a comparison study of the plug-in method with the cross-validation method shows that the plug-in method is inefficient and fails when the choice of the pilot bandwidth is wrong. Other work that recommends the CV method, or generally data driven bandwidth methods include, Izenman (1991) survey for density estimation and Chiu (1996) simulation study that compares different bandwidth estimation methods. Ouyang et al. (2006) examines asymptotic properties of the CV method and recommends the cross validation method for the MDT kernel density estimation.

Recently, cross validation methods have shown new advantages in KMWFB. The CV methods have the ability to smooth out irrelevant variables in the model as proved in Hall et al. (2004) and Li and Racine (2004). Large bandwidths for the irrelevant variables are estimated. This is consistent with the argument that the bandwidths of the irrelevant variables converge to infinity, or 1 in the discrete variables case. This advantage as shown by Li and Zhou (2005) is unique for CV among other bandwidth estimation methods. Accordingly, it has been used to established the basis of nonparametric kernel significance tests in the nonparametric regression. Compared with other nonparametric methods, KMWFB is developed more mainly because of this advantage. The kernel method with fixed bandwidth is incorporated in many software packages like STATA, SPSS, MATLAB as an advanced tools of data analysis, whilst other method are available in fewer software packages, for example R software.

However, the cross-validation methods slowing the use of the kernel method in empirical research, because they require complicated calculations. The objective function of the CV method may have several local minima (maxima) and be difficult to optimize, which may add more computation in the model. Cross validation bandwidths, additionally, show large variability when they are affected by the outliers in the model, see Park and Marron (1990). The computation cost of the CV methods is now easier with the new generations of computers and shared processing systems, like the high performance computer that used in this research, this encourages the use of the kernel method in nowadays.

To overcome the cost of numerical calculations of the CV bandwidths Botev and Kroese (2008) propose a technique that is based on Maximum Entropy method, which is quite different approach to the approach that used in the CV method. However, they argue that their new approach is data-driven method and not time consuming, and provide examples that show that Maximum Entropy method works in univariate and multivariate kernel density functions. Bors and Nasios (2009) on the other hand, proposes three methods that not numerically expensive to estimate the bandwidths in the kernel method. The methods, as introduced by Bors and Nasios (2009), very briefly, are based on ”scale space”, ”mean shift”, and ”quantum clustering”. But the work is related to engineering and industrial sciences and beyond the level of this research. The problem of the variability of bandwidths that is caused by the outliers in the model is discussed in Chen et al. (2009), and the problem of the slow convergence rate of the bandwidths is discussed by Hall and Robinson (2009).

1.2.4 Estimation of Mixed Data Types Distribution Function

The estimation of the cumulative distribution function using the kernel method is slightly more complicated than the estimation of the density function. The distribution function jumps at each ordered sample observation and at the boundaries, which make the estimators produce estimates outside the $[0, 1]$ range. The simplest estimator is the **empirical distribution function (EDF)** that takes the form $\widehat{F}_n(x^{(\cdot)}) = n^{-1}\sum_{i=1}^{n} I(X_i^{(\cdot)} \leq x^{(\cdot)})$. Where $(\cdot)$ denotes that x is either discrete or continuous. The EDF estimator of the distribution function is consistent, but exhibit a problem that it produce a step function estimate for continuous variables when the true density is smooth. Developing kernel estimators for the distribution function that overcomes this problem is considered early in the literature, as in Nadaraya (1965), Azzalini (1981), Falk (1985) and Jin and Shao (1999).

Li and Racine (2003) propose an estimator of the multivariate mixed type distribution function that uses the integrations of continuous variables kernel functions and the summation for the discrete variables kernel functions. The estimator takes the following form, Ju et al. (2009, page 4) :

$$\begin{aligned} \widehat{F}(\mathbf{x}^c, \mathbf{x}^d) &= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{G}\left(\frac{\mathbf{X}_i^c - \mathbf{x}^c}{\widehat{\mathbf{h}}_{\mathbf{x}^c}}\right)\left(\sum_{\mathbf{u}\leq\mathbf{x}^d} L\left(X_i^d, u^d, \widehat{\gamma}\right)\right)\right] \qquad (1.22) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{s=1}^{q} G\left(\frac{X_{is}^c - x_{is}^c}{\widehat{h}_s}\right)\left(\sum_{\mathbf{u}\leq\mathbf{x}^d}\prod_{r=1}^{p} l\left(X_{ip}^d, u_p^d, \widehat{\gamma}_r\right)\right)\right], \end{aligned}$$

where $G\left(y\right) = \int\limits_{-\infty}^{y} k(v)dv$. The univariate continuous distribution function is given when $q = 1$ and $p = 0$. The consistency conditions ensure that the kernel estimators of the distribution function have the following properties.

$$\sqrt{n}\left(\widehat{F}\left(\mathbf{x}^c, \mathbf{x}^d\right) - F(\mathbf{x}^c, \mathbf{x}^d)\right) \overset{d}{\rightarrow} N\left(0, F(\mathbf{x}^c, \mathbf{x}^d)(1 - F(\mathbf{x}^c, \mathbf{x}^d)\right) \qquad (1.23)$$

The bandwidth estimation methods for kernel multivariate distribution function are discussed by Sarda (1993), Altman and Leger (1995) who propose the weighted MISE to estimated the bandwidths. Bowman et al. (1998) recommend the cross validation methods for bandwidth estimation and present a comparison between different methods of smoothing parameters estimation for kernel distribution function. Hansen (2004a) proposes a plug-in estimator based on optimising AMISE. Other work includes Bowman, et al. (1998), Cheng and Peng (2002), Efromovich (2004), Kim et al. (2006), Ouyang et al. (2006), Gannoun et al. (2007), Janssen et al. (2007) and Liu and Yang (2008). Li and Zhou (2005) and Ju et al. (2009) propose the CV method to estimate the bandwidths of the MDT multivariate kernel distribution function estimator.

The asymptotic properties for kernel distribution function estimators are better than the asymptotic properties of kernel density function estimators. Where the convergence rate of the distribution function bandwidth is faster than the convergence rate of the bandwidths of the density function.

1.3 Estimation of Conditional Functions in the Mixed Data Types Framework

The core of econometric is the estimation of conditional functions. The nonparametric estimation framework is rich with methods to estimate different type of models, including: conditional density, conditional distribution, regression function and conditional hazard models. Kernel methods to estimate those functions are illustrated in this section, however, the methods to estimate the conditional hazard are discussed in Chapter 4, where is beyond the discussion in this chapter. The conditional functions are estimated directly in the kernel estimation framework without assuming a specific functional form to the relationships between the variables in the model.

1.3.1 Mixed Data Types Conditional Density Function

The kernel conditional density estimator is suggested by Rosenblatt (1969) and improved by Hyndman et al. (1996) who introduce a conditional density estimator that uses the local constant kernel regression technique. The same estimator is discussed also in Wang and Van Ryzin (1981), Chow et al. (1983), Stute (1986a,b), Hardle et al. (1988), Falk (1993) and Hyndman et al. (1996). The methods to estimate the bandwidths for the conditional density are discussed in Turlach (1993) and Bashtannyk and Hyndman (2001). Racine et al. (2004) and Hall et al. (2004) develop the mixed type conditional density function estimator for the densities with categorical and continuous variables. Racine et al. (2004) recommend the Modified Aitchison and Aitken kernel function in Tables 1.2 and 1.3 for the categorical unordered and categorical ordered variables respectively.

Let $\mathbf{y}$ be the vector of the values of mixed type dependent variables, where $\mathbf{y} = [y_1^c, y_2^c, ..., y_{q_y}^c,\ y_1^d, y_2^d, ..., y_{p_y}^d]$, and $\mathbf{x}$ be the vector of the values of the independent variables $\mathbf{x} = [x_1^c, x_2^c, ..., x_{q_x}^c,\ x_1^d, x_2^d, ..., x_{p_x}^d]$. Then, for simplicity, denote the variables in the joint density by $\mathbf{z}^c$ and $\mathbf{z}^d$, then $\mathbf{z}^c = [y_1^c, y_2^c, ..., y_{q_y}^c,\ x_1^c, x_2^c, ..., x_{q_x}^c]$ and $\mathbf{z}^d = [y_1^d, y_2^d, ..., y_{q_y}^d,$ $x_1^d, x_2^d, ..., x_{q_x}^d]$. The kernel conditional density estimator takes the following form, rearranged from Li and Racine (2007, Chapter 4):

$$
\begin{aligned}
\widehat{f}_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{y}\,|\mathbf{x}\right) &= \frac{\widehat{f}_{\mathbf{Y},\mathbf{X}}(\mathbf{y},\mathbf{x})}{\widehat{f}_{\mathbf{X}}(\mathbf{x})} \\
&= \frac{n^{-1}\sum\limits_{i=1}^{n}\prod\limits_{s=1}^{q} h_s^{-1} w\left(\frac{Z_{is}^c - z_s^c}{h_s}\right)\cdot\prod\limits_{r=1}^{p} l\left(Z_{ir}^d, z_r^d, \gamma_r\right)}{n^{-1}\sum\limits_{i=1}^{n}\prod\limits_{s=1}^{q_x} h_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{h_s}\right)\cdot\prod\limits_{r=1}^{p_x} l\left(X_{ir}^d, x_r^d, \gamma_r\right)},
\end{aligned}
\tag{1.24}
$$

where, $z^{(\cdot)}$ refers to any $y^{(\cdot)}$ or $x^{(\cdot)}$ in the joint density function. q is the number of continuous variables, p is the number of discrete variables, in the model and in the joint density function respectively. q_x is the number of independent continuous and p_x is the number of discrete independent variables and so they are the number of continuous and discrete variables in the marginal density function respectively. $w(\cdot)$ is a univariate continuous variable kernel function, see Table 1.1. $l(\cdot)$ is a discrete univariate Aitchison and Aitken kernel function, from Table 1.2 if the variable is unordered and from Table 1.3 if the variable is ordered.

The number of the continuous dependent variables is $q_y = q - q_x$ and the number of discrete dependent variables is $p_y = p - p_x$. For example, when $q_y = 1$ and $p_y = 0$, the conditional density model includes one continuous dependent variable. When $q_y = 0$ and $p_y = 1$ the conditional density includes only one dependent discrete variable. The kernel conditional density model could then be considered as the kernel nonparametric equivalent of the discrete choice parametric model. Hall et al. (2004) recommend the cross-validation method to estimate the bandwidths of the conditional density as illustrated below.

The LSCV bandwidths for the conditional density with mixed variables type is based on minimizing the sample analogue of the weighted integrated square error I_n given by the following formula:

$$I_n = \left\{ \widehat{f}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}\,|\mathbf{x}) - f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}\,|\mathbf{x}) \right\}^2 \widehat{f}_{\mathbf{X}}(\mathbf{x}).M(\mathbf{x})d\mathbf{x}d\mathbf{y}, \tag{1.25}$$

where $M(\mathbf{x})$ is a nonnegative weighting function.

The MLCV bandwidths on the other hand are estimated by maximizing the following function:

$$L = \sum_{i=1}^{n} \log \widehat{f}_{\mathbf{Y}|\mathbf{X}-i}(\mathbf{Y}_i\,|\mathbf{X}_i). \tag{1.26}$$

The bandwidths satisfy $h_s \geq 0$ for continuous variables and $0 \leq \gamma_r \leq (c_r - 1)/c_r$ discrete variables. Racine et al. (2004) show that the bandwidths that are estimated for the conditional density function with MDT have the following asymptotic properties:

The bandwidth selected by CV converges to the optimal parameters at a rate $O(n^{-1/q+5})$ for the bandwidths of the continuous variables and $O(n^{-2/q+5})$ for the discrete variables.

The bandwidths that are estimated by the maximum likelihood cross-validation method may oversmooth the kernel density if the continuous variables in the model are drawn from fat-tail distributions. But the maximum likelihood cross-validation is more useful than the least squares cross validation method because it requires less numerical calculations and is easier to apply in empirical research.

1.3.2 Conditional Distribution Function

Kernel estimation of **conditional distribution function CDF** for continuous variable is established by Nadaraya (1964). Azzalini (1981) uses two different ways to drive the conditional distributions estimators: first by using direct integration of the kernel function of the dependent variable, and second by estimating the quantiles. Those two methods are discussed in Falk (1984), Chu and Marron (1991), Fan (1993), Fan, Yao and Tong (1996) and Yu and Jones (1997) and Yu and Jones (1998).

The estimation of the conditional distribution function is illustrated in Li and Racine (2007). The conditional distribution function for one continuous dependent variable takes the form:

$$F_{Y|\mathbf{X}}(y^c|\mathbf{x}) = \int_{-\infty}^{y^c} \frac{f_{Y,\mathbf{X}}(u,\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} du, \tag{1.27}$$

which is estimated by

$$\widehat{F}_{Y|\mathbf{X}}(y^c|\mathbf{x}) = \int_{-\infty}^{y^c} \frac{\widehat{f}_{Y,\mathbf{X}}(u,\mathbf{x})}{\widehat{f}_{\mathbf{X}}(\mathbf{x})} du, \tag{1.28}$$

where $\widehat{f}_{Y,\mathbf{X}}(y^c,\mathbf{x})$ and $\widehat{f}_{\mathbf{X}}(\mathbf{x})$ are the kernel estimators of the joint and marginal density functions respectively.

There are two approaches for estimating the conditional distribution functions. The first approach uses Nadarya-Watson kernel regression, the local constant nonparametric regression, to estimate the distribution function using an indicator function $I\left(Y_i^c \leq y^c\right)$. This estimator is proposed by Hall et al. (1999) and is known as the **conditional mean CDF** estimator, and takes the following form in the mixed data types case, see Li and Racine (2007, page 194):

$$\widehat{F}^{(1)}_{Y|\mathbf{X}}(y^c|\mathbf{x}) = \frac{n^{-1}\sum_{i=1}^{n} I\left(Y_i^c \leq y^c\right) \prod_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right) \prod_{r=1}^{p} l\left(X_{ir}^d, x_r^d, \widehat{\gamma}_r\right)}{n^{-1}\sum_{i=1}^{n}\prod_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right) \prod_{r=1}^{p} l\left(X_{ir}^d, x_r^d, \widehat{\gamma}_r\right)}. \tag{1.29}$$

The second estimator substitutes the indicator function $1(Y_i^c \leq y^c)$ by an integrated kernel function of y^c, $G(y^c) = \int_{-\infty}^{y^c} k(y^c)dy^c$, as follows, see Racine (2008, page 27) and Li and Racine (2007):

$$\widehat{F}^{(2)}_{Y|\mathbf{X}}(y^c|\mathbf{x}) = \frac{n^{-1}\sum_{i=1}^{n} G\left(\frac{Y_i^c - y^c}{\widehat{h}_0}\right) \prod_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right) \prod_{r=1}^{p} l\left(X_{ir}^d, x_r^d, \widehat{\gamma}_r\right)}{n^{-1}\sum_{i=1}^{n}\prod_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right) \prod_{r=1}^{p} l\left(X_{ir}^d, x_r^d, \widehat{\gamma}_r\right)}. \tag{1.30}$$

The consistency condition for the conditional distribution function are that as $n \to \infty$, $h_s \to 0$, $nh_1 ... h_m \to \infty$.

The estimation of the smoothing parameters for the conditional distribution function is discussed in Hyndman et al. (1996) and Yu et al. (2008). Hall et al. (1999) proposes a bootstrap method based on AIC and point to the difficulties on estimating the bandwidths of conditional densities. A cross validation method for the conditional distribution function is proposed by Bashtannyk and Hyndman (2001) and Hyndman and Yao (2002). Hansen (2004b), in an unpublished manuscript, provides a review of the literature of bandwidth estimation methods for the kernel conditional densities and conditional distribution function. Hall et al. (2004) provides the methods for estimating the bandwidths for the conditional distribution function, which is similar to the methods of estimating the bandwidths for the conditional densities. However, the optimum bandwidths of the conditional distribution as shown by Hall et al. (2004), are slightly different than the optimum bandwidths of the conditional density function. Hence, researchers should be careful when applying the bandwidths that are estimated for one function in another.

1.3.3 Nonparametric Regression

Kernel nonparametric regression estimators are more popular in the literature than kernel conditional density and conditional distribution function estimators. Applications of kernel nonparametric regression are available more and discussed thoroughly. Kernel regression is introduced separately by Nadaraya (1965) and Watson (1964), and extended by Priestley and Chao (1972), Priestley and Chao (1972) and Benedetti (1977) among many others. In this research, kernel regression is used in Chapter 4 in the sub-survival functions in the conditional hazard estimator. This section illustrates the development of kernel regression technique with MDT regressors.

The conditional moment function takes the form:

$$Y = E\left(Y \mid \mathbf{X} = \mathbf{x}\right) + u = m(\mathbf{x}) + u, \tag{1.31}$$

where $m(\mathbf{x})$ is the optimal prediction of Y given $\mathbf{X}$. Pagan and Ullah (1999), define nonparametric regression approach as a method to "approximates the function arbitrarily". This incorporates trying to estimate the densities in the conditional expectation function directly as follows:

$$E\left(Y \mid \mathbf{X} = \mathbf{x}\right) = m(\mathbf{x}) = \int \frac{y.f_{Y,\mathbf{X}}(y, \mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} dy, \tag{1.32}$$

which is estimated as follows:

$$\widehat{m}(\mathbf{x}) = \int \frac{y\widehat{f}_{Y,\mathbf{X}}(y, \mathbf{x})}{\widehat{f}_{\mathbf{X}}(\mathbf{x})} dy. \tag{1.33}$$

The simplest kernel regression estimator is the **local constant regression**, developed independently by Nadaraya (1965) and Watson (1964) and widely known as **Nadarya-Watson regression**.

The nonparametric kernel regression estimator of y on a single continuous independent variable x^c is given as follows:

$$\begin{aligned}\widehat{m}(x) &= \min_{b_0} \sum_{i=1}^{n} \{Y_i - b_0\}^2 \times k\left(\frac{X_i - x}{h}\right) \\ &= \frac{\sum\limits_{i=1}^{n} Y_i w\left(\frac{X_i - x}{h}\right)}{\sum\limits_{i=1}^{n} w\left(\frac{X_i - x}{h}\right)}, \end{aligned} \tag{1.34}$$

which is considered as a local weighted average of the dependent variable. The local weights, which are known as the Nadarya-Watson weights, are given by the kernel function of the independent variables in the model $\sum\limits_{i=1}^{n} \frac{w\left(\frac{X_i - x}{h}\right)}{\sum\limits_{i=1}^{n} w\left(\frac{X_i - x}{h}\right)} = 1$. The kernel regression of y on MDT independent variables is given by using the product of the kernel functions of the independent variables in the Nadarya-Watson weights as follows:

$$\widehat{m}(\mathbf{x}) = \frac{n^{-1} \sum\limits_{i=1}^{n} Y_i \prod\limits_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right) \prod\limits_{r=1}^{p} l\left(X_{ir}^d, x_r^d, \widehat{\gamma}_r\right)}{n^{-1} \sum\limits_{i=1}^{n} \prod\limits_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right) \prod\limits_{r=1}^{p} l\left(X_{ir}^d, x_r^d, \widehat{\gamma}_r\right)}. \tag{1.35}$$

Or simply

$$\widehat{m}(\mathbf{x}) = \sum_{i=1}^{n} Y_i \mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}}), \tag{1.36}$$

where $\mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}})$ are the **Nadaraya-Watson weights** that given by

$$\mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}}) = \frac{n^{-1} \prod\limits_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right) \prod\limits_{r=1}^{p} l\left(X_{ir}^d, x_r^d, \widehat{\gamma}_r\right)}{n^{-1} \sum\limits_{i=1}^{n} \prod\limits_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{\widehat{h}_s}\right) \prod\limits_{r=1}^{p} l\left(X_{ir}^d, x_r^d, \widehat{\gamma}_r\right)}. \tag{1.37}$$

The local constant kernel estimator is highly discussed in the literature, among those who improved this estimator are Priestley and Chao (1972), Stone (1977, 1982), Benedetti (1977), Clark (1977), Gasser and Müller (1979), Gasser and Müller (1984), Gasser et al. (1985), Cheng and Lin (1981), Ahmad and Lin (1984), Muller and Stadtmuller (1987), Bierens (1987) and Chu and Marron (1991).

The local constant regression is extended to a **local linear** and to a **local polynomial** kernel regression by Fan et al. (1992), Fan (1993) and Fan, Hall, Martin and Patil (1996), using the method of Stone (1977). The local polynomial function is a Taylor expansion of the local linear function and improves the consistency properties in the nonparametric regression by substituting the simple Nadaraya-Watson weights by a sequence of probability weighting functions. The local polynomial regression attracts more attention in the literature especially when the objective is to reduce the bias in the kernel estimates. The local constant regression though is a special case of the local polynomial regression. The general form for the local polynomial nonparametric kernel regression with one independent continuous variable is, see Li and Racine (2007, page 85):

$$\begin{aligned}\widehat{m}(x) &= \min_{b_0,b_1,\ldots,b_d}\sum_{i=1}^{n}\left\{Y_i - b_0 - b_1(X_i-x) - \ldots. - b_d(X_i-x)^d\right\}^2 \\ &\quad \times k\left(\frac{X_i-x}{h}\right), \qquad (1.38)\end{aligned}$$

which is known as the local polynomial regression function of order d. When $d = 0$, the formula reduces to the local constant regression, Li and Racine (2007, page 61). When $d = 1$ the model has a first order polynomial, which is known as the local linear regression and has the form:

$$\widehat{m}(x) = \min_{b_0,b_1}\sum_{i=1}^{n}\left\{Y_i - b_0 - b_1(X_i-x)\right\}^2 \times k\left(\frac{X_i-x}{h}\right). \qquad (1.39)$$

Härdle et al. (2004, pages 94-96) shows the following weighted least squares formula for the local polynomial kernel regression. The same formula is presented also in Cleveland (1979), Hastie and Loader (1993), Mays (1995), Mays et al. (2001) and Racine (2008). To make the local polynomial kernel regression written in weighted least squares form: first, sort the difference between X_i and x and the values of the dependent variable Y_i in matrices as follows:

$$\mathbf{X} = \left[c, \widetilde{X}\right] = \begin{bmatrix} 1 & (X_1-x) & (X_1-x)^2 & \cdots & (X_1-x)^d \\ 1 & (X_2-x) & (X_2-x)^2 & \cdots & (X_2-x)^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n-x) & (X_n-x)^2 & \cdots & (X_n-x)^d \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix},$$

$\mathbf{X}$ has an order $n \times (1+d)$, for local constant model $\mathbf{X}$ is $n \times 1$, a column of ones a. In the local linear regression $\mathbf{X}$ has order $n \times 2$. Second, define the diagonal matrix $\mathbf{W}$ that includes the kernel weights

$$\mathbf{W} = h^{-1} diag\left[k(\tfrac{X_1-x}{h}), \quad k(\tfrac{X_2-x}{h}), \quad \cdots \quad k(\tfrac{X_n-x}{h})\right].$$

Define also the matrix $\mathbf{B}$ and the vector $\mathbf{S}$ as follows:

$$\begin{aligned}
\mathbf{B} &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1} \\
&= \begin{bmatrix}
\sum\limits_{i=1}^{n} k(\frac{X_i-x}{h}) & \sum\limits_{i=1}^{n} (X_i-x)\, k(\frac{X_i-x}{h}) & \cdots & \sum\limits_{i=1}^{n} (X_i-x)^d\, k(\frac{X_i-x}{h}) \\
\sum\limits_{i=1}^{n} (X_i-x)\, k(\frac{X_i-x}{h}) & \sum\limits_{i=1}^{n} (X_i x)^2\, k(\frac{X_i-x}{h}) & \cdots & \sum\limits_{i=1}^{n} (X_i-x)^{d+1}\, k(\frac{X_i-x}{h}) \\
\vdots & \vdots & \ddots & \vdots \\
\sum\limits_{i=1}^{n} (X_i-x)^d\, k(\frac{X_i-x}{h}) & \sum\limits_{i=1}^{n} (X_i-x)^{d+1}\, k(\frac{X_i-x}{h}) & \cdots & \sum\limits_{i=1}^{n} (X_i-x)^{2d}\, k(\frac{X_i-x}{h})
\end{bmatrix}^{-1} \\
&= \begin{bmatrix}
\beta_0 & \beta_1 & \cdots & \beta_d \\
\beta_1 & \beta_2 & \cdots & \beta_{d+1} \\
\vdots & \vdots & \ddots & \vdots \\
\beta_d & \beta_{d+1} & \cdots & \beta_{2d}
\end{bmatrix}, \qquad (1.40)
\end{aligned}$$

and

$$\mathbf{S} = \mathbf{X}^T\mathbf{W}\mathbf{Y} = h^{-1} \begin{bmatrix}
\sum\limits_{i=1}^{n} k(\frac{X_i-x}{h}) Y_i \\
\sum\limits_{i=1}^{n} (X_i-x)\, k(\frac{X_i-x}{h}) Y_i \\
\sum\limits_{i=1}^{n} (X_i-x)^2\, k(\frac{X_i-x}{h}) Y_i \\
\vdots \\
\sum\limits_{i=1}^{n} (X_i-x)^d\, k(\frac{X_i-x}{h}) Y_i
\end{bmatrix} = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ \vdots \\ s_d \end{bmatrix} \qquad (1.41)$$

$$(1.42)$$

and

$$\mathbf{e}^T = [1, 0, 0, ...0]\,. \qquad (1.43)$$

where e has value 1 at the first element and zero elsewhere.

When the model includes more than one independent variable the distances in $\mathbf{X}$, $\mathbf{B}$ and $\mathbf{S}$ are replaced by the euclidean distances between the vectors $\mathbf{X}_i$ and x. $\mathbf{W}$ includes the kernel products $\prod\limits_{s=1}^{q} h_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{h_s}\right) \prod\limits_{r=1}^{p} l\left(X_{ir}^d, x_r^d, \gamma_r\right)$. Then the local polynomial regression function becomes

$$\widehat{m}(\mathbf{x}) = \mathbf{e}^T\left(\mathbf{X}^T\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\mathbf{W}}\mathbf{Y} = \mathbf{e^T}\mathbf{B}\mathbf{S} = \sum_{j=0}^{d} \beta_j s_j, \qquad (1.44)$$

which is the regression estimate at the point x, and $\widehat{W}$ is the diagonal weights matrix that is generated form the kernel functions and the estimated bandwidths. This shows that the local polynomial kernel regression is not easy to compute because it includes many smoothed terms.

The precision of the kernel regression depends on the choice of the kernel functions in the model, higher order kernels functions reduce the bias in the model, but on the other hand are harder to compute. It also depends of the choice of the bandwidth estimation method, CV method is recommended due to its nice properties of the estimated bandwidths. The higher order local polynomial regression also improves the precision of the model, but increases the computation cost.

Estimation of the bandwidths of the kernel nonparametric regression is discussed in Rice (1984), Hardle et al. (1992), Muller and Stadtmuller (1987), Azzalini et al. (1989), Vieu (1991), del Río (1996), Herrmann (1997) and Ziegler (2002). The Plug-In method and Rule of Thumb are inefficient compared with the automatic data driven bandwidth estimation methods, similar to the case of density and conditional density estimation. Lee and Solo (1999) compare between the bandwidth estimation methods for local linear regression using simulation method, the Rule-of-Thumb method and the Plug-in method show inefficiency compared with LSCV method. A similar comparison study is presented by Loader (1999) who demonstrates the same result for the Plug-in method and CV methods.

The LSCV method estimates the optimum bandwidth by optimizing the following objective function, Li and Racine (2007, page 138):

$$CV_{ls}(h,\gamma) = \sum_{i=1}^{n}(Y_i - \widehat{m}_{-i}(X_i))^2 M(X_i) \tag{1.45}$$

where $\widehat{m}_{-i}(X_i)$ is the leave-one-out kernel estimator for $m(x)$ and $M(X_i)$ is a weighting function.

1.4 Advanced Kernel Models

Owen (2001) suggests a method of inference of parametric model using nonparametric likelihood ratios, defined as the nonparametric likelihood method, that is maximized by the empirical distribution function. The EL method reassigns weights, p_i, to the sample observations, instead of the Uniform weights n^{-1}, in an attempt to improve the nonparametric likelihood function $L(F) = \prod_{i=1}^{n} p_i$ under the restriction that $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. The method uses a profile likelihood function of the parameter that being estimated. Adopting the same procedure in nonparametric kernel estimation is attempted in two different approaches. First, the **weighted kernel density** estimator, also known as the **globally weighted kernel density** and the **re-weighted kernel density**, that is suggested by Hall and Turlach (1999) and Hazelton and Turlach (2007), and discussed also by Wang and Wang (2007). The second is defined as the constrained kernel density is a kernel density and developed by Chen (1997), is considered a stronger empirical likelihood base kernel density estimator that the weighted density.

1.4.1 Weighted Kernel Density Estimators

The weighted kernel density is estimated by the following formula:

$$\widehat{f}_p^*(x^c) = \frac{1}{\widehat{h}}\sum_{i=1}^{n} p_i k\left(\frac{X_i^c - x^c}{\widehat{h}}\right), \tag{1.46}$$

where: $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$ are defined by Hall and Turlach (1999) as **global weights**, where they are estimated from $\{X_i\}_{i=1}^{n}$ and do not depend on x^c and the distance $X_i^c - x^c$. Since the weights that are given by the kernel function are defined as the local weights, where the observations are weighted in a distance controlled by the value of the bandwidth. Hazelton and Turlach (2007) referred to this technique as the reweighting of the kernel density. The properties of the weighted kernel density estimator are discussed by Hall and Turlach (1999). The weighted kernel density estimator has less bias than the traditional kernel density estimator.

Wang and Wang (2007) focus on the method of estimating of the optimum bandwidths for the weighted kernel density, which is shown to be not substantially different from the optimum bandwidths of the traditional kernel density estimator. Hall and Turlach (1999) and Hazelton and Turlach (2007) on the other hand, focus on the choice of the optimum global weights for the estimator. Both papers examine the kernel density estimator that suggested by Marron and Padgett (1987) as a special case of the weighted kernel density estimator.

The predicted global weights are defined by Hall and Turlach (1999) as the "bias annihilating global weights". Hazelton and Turlach (2007) note that the weighted kernel density estimation framework is useful to estimate models of mixed distributions. Marron and Padgett (1987) and Wang and Wang (2007) use the approach for right censored observations. The theoretical validation of the global weighting idea of the kernel density estimators is built on the convolution the of the kernel functions with other continuous functions. Hazelton and Turlach (2007, page 3060) in Theorem 1 define the optimum global weights as function on unknown densities $g(\cdot)$ and $f(\cdot)$ as $p_{0,i} = \frac{1}{n}\frac{g\left(X_i^c\right)}{f\left(X_i^c\right)}$. Hall and Turlach (1999) illustrate that "bias annihilating global weights are positive and satisfy $E\left[\sum_{i=1}^{n} p_i\right] = 1$".

A data-driven method is recommended to estimate the optimum global weights using a similar approach to that used in the estimation of the bandwidths of the density. The restrictions that the weights must be non-negative and have sum equals 1 must be added to the optimization formula. The global weights can only be estimated under known bandwidths, either Plug-in or cross validation bandwidths. The computation of the objective function of the global weights is very hard compared to the computations to estimate the bandwidths. Hall and Turlach (1999) suggests a method based on minimizing the ISE, similar to the LSCV in bandwidth estimation, but with the restrictions regarding the global weights added to the function.

The Hall and Turlach (1999) method is not easy to use empirically. Hall and Turlach (1999) and Hazelton and Turlach (2007) show that the data-driven global weights are almost impossible computationally when the sample size exceeds 500 observations. The discussion of the weighted density is presented on the context of univariate continuous variable densities, and the application of this method to discrete univariate and conditional densities appears not to have been attempted. The convolution based global weights makes this method unobvious to use for the univariate discrete variable case as well as in the case of mixed variables type densities.

1.4.2 Constrained Kernel Density Estimation

Chen (1997) suggests the following kernel density estimator:

$$\widehat{f}^*_{el}(x^c) = \frac{1}{\widehat{h}}\sum_{i=1}^{n} p_i k\left(\frac{X_i^c - x^c}{\widehat{h}}\right), \tag{1.47}$$

where: $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. p_i are estimated by maximizing a multinomial likelihood function $\prod_{i=1}^{n} p_i$, but additional information is needed in the model. Chen (1997) expresses this information in the moment condition $E_X\{g_l(X^c)\}$, where g_l, $l = 1, \cdots, m$ are real functions. The density function is estimated with the conditions that $\sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} p_i g_l(X_i^c) = 0$ for $l = 1, \cdots, m$.

The optimum weights are shown by Chen (1997, page 49) have the following form:

$$p_i = n^{-1}\{1 + \lambda^T \mathbf{g}(X_i^c)\}, \tag{1.48}$$

where $\lambda = (\lambda_1, \lambda_2, ..., \lambda_m)^T$ are the Lagrange multipliers and $\mathbf{g}(X_i^c) = (g_1(X_i^c), g_2(X_i^c), \cdots, g_m(X_i^c))^T$.

The additional information can be constructed from the data, Chen (1997) suggests the first moment, $E(X^c - \mu_0)$ for the unknown mean μ_0, and shows that the empirical likelihood density estimator in this case takes the form, Chen (1997, page 50, Eq (7)) :

$$\widehat{f}^*_{el}(x^c) = \frac{1}{n\widehat{h}}\sum_{i=1}^{n} \frac{1}{1 + \lambda(X_i^c - \mu_0)} k\left(\frac{X_i^c - x^c}{\widehat{h}}\right). \tag{1.49}$$

Hall and Presnell (1999) extend this method to use general constrains, and shows that additional constraints can be used, such as the unimodality constrain which is adopted in Hall and Huang (2002) and Hall and Kang (2005).

1.5 Comparison Between Mixed Data Types Kernel Estimation Methods and Parametric Estimation Methods

The nonparametric kernel technique is a revolutionary method in econometric estimation, has so far had application in a few types of models. The kernel method is vague in models with random effects and panel data and on handling the endogenity problem in the data. The performance of the significance tests for kernel methods is difficult and time consuming in empirical research. The results of kernel estimators are presented side-by-side with the results of the corresponding parametric model. There are additional sources of differences that we distinguished during the work in this thesis. On the following a highlight of some of those differences as they appeared to us.

First, nonparametric kernel estimators do not produce coefficients like their parametric counterparts. The only parameters that are produced are the smoothing parameters, the bandwidths, that are used to predict the object that being estimated. This key difference has many implications in interpreting and in examining nonparametric estimators results. The coefficients in the parametric models, particularly the sign of the coefficients, reflect the type of the effect that independent variables make in the model, whether negative or positive. The size of the effect, the elasticities and many other issue related to the type of the model that being estimated are also known from the estimated coefficients. In the kernel nonparametric estimation, the bandwidths do not carry the same information and do not have the same rules. Researchers have to compute the predicted values of the function at chosen values of the independent variables to know the shape of the function and the relationships in the estimated kernel object. Hence, working with kernel estimators needs more skills and knowledge about the function that is being estimated than in parametric estimators.

The consequences of these differences in empirical research are that: using tables to compare between the results of parametric and kernel nonparametric estimator becomes unrevealing. Graphs and plots became more effective and emphasise, but must be handled carefully. Calculating the predicted values can be misleading if not conducted carefully. Pointwise sample predictions are usually confusing in plots, since researchers have to set a few independent variables fixed at certain values as appropriate to the discussion and the objectives in the research. Usually the graphical presentation is only successful after examining a large number of plots and becomes more difficult if the model includes many variables and there are a large number of cells.

Second, the nonparametric kernel estimation method applies large amount of numerical calculations to estimate the bandwidths and to predict the function in the model. This makes the application of the kernel method difficult in studies with large sample sizes or that include many variables. Computing of the differences and weighting them by the kernel function for each pair of observations in the sample is a tedious process. For the estimators that include many smoothed objects, like the sub-survival (sub-distribution) functions in the conditional kernel hazard estimators in Chapter 3, the cost of the computation process increases substantially. Consequently, the estimation of the optimum bandwidths using data-driven method becomes a very hard, since the pointwise sample predictions are estimated repeatedly to compute the objective function at each iteration until the optimum bandwidths are achieved.

?? Third, in parametric estimation the categorical independent variables are added as dummy variables, with little distinction between between unordered and ordered variables. A categorical independent variable with 4 choices (discrete values) is represented in a model that estimated parametrically by 3 dummy variables if the intercept is included. In the MDT estimation of the same model the same categorical variable is smoothed directly without converting it to dummy variables. Aitchison and Aitken (1976), illustrate the advantages of discrete kernel function in this context using the cluster analysis. Consider a discrete variable, X^d, that specified as dummy variables in MDT kernel estimators, with all the dummies smoothed by a same type of discrete kernel function chosen from Table 1.2 or Table 1.3. Aitchison and Aitken (1976) show that the product of the kernel functions of the dummy variables allocate the same number of weights that a single discrete kernel function allocates to X^d. Then the optimum bandwidths for the kernel functions of the dummy variables will be estimated with paying high unnecessary cost to smooth the data, where computing the weights by kernels product apply or more procedures than computing the weights by a single kernel function.

Hence, the number of covariates, which is the independent variables with the continuous variables transformed to define a functional form and the discrete variables converted to dummy variables, in the parametric estimator will be different than in the MDT estimator. The covariates set for parametrically estimating the model will be longer if many continuous variable are included with transformed version to define a quadratic or a linear polynomial form, and one more discrete variables have many possible outcomes. For example, to capture the effect of calendar months in a parametrically specified model 11 dummy variables are needed, to estimate the same model nonparmetrically a singe discrete ordered variable that take integer values from 1 to 12 captures the effect. Accordingly, number of covariates that are needed to perform the MDT kernel estimation is likely to be lower depending on the type of the independent variables and the object that being estimated as well as on the parametric specification of the model. The number of variables becomes different if the parametric model defines functional form for the continuous variables, like a quadratic or a linear polynomial form, or one or more discrete variable has more than two possible outcomes.

Finally, the estimation of elasticities for kernel regression and kernel conditional density functions is not sufficiently developed. Pagan and Ullah (1999) illustrates a number of numerical methods and kernel based methods for estimation of marginal effects in nonparametric regression. Additional methods are available in Gasser and Müller (1984), Jiang and Doksum (2003), Abdous et al. (2003) and Lee and Lee (2008). The estimation of the marginal effects and the derivatives of kernel conditional functions may become very complicated, for example the marginal effects of the conditional hazard estimator in Chapter 4. This in one hand restrict the use and the interpretation of the results of kernel estimators, but on the other hand opens a wide field of further research in kernel method estimation framework.

1.6 Research Objectives

The objective of this PhD thesis is to apply empirically some of the recently developed kernel nonparametric estimators for MDT to estimate models of discrete dependent variables, and compare the results with the regular parametric link function estimators using labour market data from the UK. The thesis comprises two applications using real data and one using simulated data. The research is faced with a number of difficulties.

First, problems like unobserved heterogeneity in the data and the curse of dimensionality problem in the nonparametric estimators, restrict the choice of independent variables in the empirical models. Second, the computation cost of the nonparametric estimators, and the calculation cost of the nonparametric smoothing tests, restrict our choice of variables in each example. We refer to the literature to choose the variables in the parametric models, the variables that are commonly used are likely to be found relevant in the nonparametric estimators. Some difficulties in our work are generated by the size of the samples that are used in each application. This problem has been overcome by using the **high performance computer (HPC)** at the University of Manchester.

The research compares critically between the results of the kernel estimator of the DCF and its corresponding parametric discrete dependent variable model in each example. To make the comparison between parametric and nonparametric estimation techniques clear, under the above restrictions, the thesis estimates models that are well known and thoroughly discussed in labour economics literature. The first empirical example focuses on female labour force participation decision using UK data, and estimates a model using the multinomial logit parametric technique and compares it with the kernel conditional density nonparametric estimator. The second empirical example estimates a discrete time conditional hazard rate model for job vacancies in Lancashire area in the UK between the period from March 1988 to June 1992 and compares the results with the results of the conventional mixed proportional hazard estimation framework.

The literature in labour economics provides sufficient information to examine the result of each estimation technique. The vast literature about female labour force participation decision is informative about the relationship between the propensity of female to work with fertility, marital status, household characteristics and female eduction level. The discussion about the econometric estimation technique that capture those relationship correctly is vast. There is a degree for dis-satisfaction about the results of the parametric models in empirical research in labour economics. Thus, we refer to the theory of female labour market supply to assess the results that delivered by each estimation technique in the first example.

In the second example, which attempts to estimate a conditional hazard model for job vacancies, the **random matching theory (RMT)** of the job-market is used to assess the estimated results of each technique. The random matching theory explains the relationship between the number of job-seeker and the number of job-vacancies with the matching rate, the average number of filled job-vacancies per unit of time, using a matching function of specific form. The RMT defines the job-market tightness as the ratio between the number of job-vacancies and the number of job-seekers, which directly influence in the speed at which the job-vacancies are filled. So, the RMT is informative about the relationship of the hazard rate of the job-vacancies with the job-market tightness, because the transformation of the matching function to a conditional hazard rate is easy.

The parametric model is estimated first in each example. The approach in this research is to find the "best fitting parametric model" first before attempting the nonparametric estimation. Which means that a model selection criteria is used in each example. All the issues that related to the reshaping of the dataset in the parametric and the nonparametric estimation frameworks are illustrated sufficiently in the thesis. Generally a correctly specified parametric model is attracted more than a nonparametric estimator, where parametric estimation is easier. However, reaching a correctly specified parametric model is difficult, particularly for DCF, where many objects like the link function, the distribution of error term, the functional form of the heterogeneity function, etc, are all need to be correctly specified in the model. Then the best fitting parametric model seems as a difficult form to be achieved. Our results show that it is difficult to define which parametric model is best fitting. In all examples we claim that the best fitting model is reached when some of the restrictive parametric assumption are relaxed.

The original contributions of this thesis to knowledge are that: (i) It applies a new, and powerful, estimation technique for British labour market data for the first time. (ii) Develops and introduce a new discrete time kernel hazard estimator in econometrics that may have many useful applications in duration models. (iii) Invents a **data generating process (DGP)**, based on the conditional proportional hazard estimation framework, that is designed in continuous time but discretized for grouped time (discrete time) hazard rate. The DGP has the advantage of controlling the random right censoring problem in the simulation. The examination of the results of the independent variables is allowed for both the continuous time and the discrete time hazard rate estimates. A competing risks right censoring process is simulated in the study to control for the proportion of the right censored observations. The DGP is introduced from the medical survival model literature, and it has been designed to examine our new developed discrete time kernel hazard estimators. (iv) The thesis presents new quality empirical results in the examples, which could be useful in labour economics decision making and future research. The results of the nonparametric estimators generally match the suggestions of the theories of labour economics better than the results of the parametric models.

1.7 Computing the Models

Kernel estimation methods for conditional functions are available in very few econometric software packages. The kernel estimators of conditional models with MDT, up to our knowledge, are available only in the software package "**R**". Other software packages that used in econometrics like Matlab and Gauss include codes that smooth continuous variables only. Matlab includes codes for univariate and multivariate kernel density estimation and kernel nonparametric regression. The codes that provided by Yi Cao allow Matlab users to estimate different objects including local constant and local linear kernel regression. In addition to those codes the free **Kernel Density Estimation Toolbox for MATLAB** offers different types of estimators. However, the codes that are available in Matlab estimate unconditional density and kernel regression function only. We have not been successful to find codes that estimate conditional density and codes that smooth discrete variables in Matlab.

Similarly the software Mathematica and Gauss include codes of kernel smoothing of continuous variables only as that available in Matlab. But they seem to provide less number of kernel estimators. The software SAS, on the other hand, include a procedure called **Kernel Density Estimation Procedure**, which performs the kernel density and kernel regression estimators. The same type of estimators are available in packages in STATA, Limdep, Nlogit and StatsDirect in addition to many other software. Those same software, however, provide estimation of many types of parametric discrete dependent variable models with different link functions. However, the kernel estimators that are included in those software can just estimate unconditional density and kernel regression only, with no capability to smooth discrete variables.

R software package, on the other hand, include many packages for kernel density estimation and kernel smoothing. For example the packages **kerdiest** (nonparametric kernel estimation of the distribution function), **NP** (nonparametric kernel smoothing methods for mixed data types), **npRmpi** (parallel nonparametric kernel smoothing methods for mixed data types) and **SM** (smoothing methods for nonparametric regression and density estimation). The smoothing methods of discrete variables is available in **NP** and **npRmpi** packages of Jeffrey S. Racine and Tristen Hayfield. The **NP** is the earlier, the first version it was published in November 2006, and it was the only package that allows to smooth the MDT in R when we started the work in this thesis.

The advantages of R software are that it is flexible, easy and free. Those advantages facilitate the work in all chapters in this thesis, where they allow us to write our own codes for kernel hazard estimators in addition to the codes of bandwidth estimation in Chapter 4. The parametric models are estimated in R and in STATA software jointly. The former include commands that allow data to be imported directly from STATA. The computation cost of the kernel estimators in this research is high so much so that there are some codes of density estimates executed in weeks. The models in the simulation study in Chapter 3 are simplified to reduce the computation time. Without the High Performance Computer of the University of Manchester some of the models and the simulations in this thesis could be impossible.

The software package R is used to estimate the multinomial discrete choice model in Chapter 2 and the homogeneous discrete time proportional hazard model in Chapter 3. To estimate the parametric discrete time conditional hazard using the mixed proportional hazard model we use STATA software. Where STATA provides more choices for the distribution function of the unobserved heterogeneity than other software like R, Limdep and Nlogit. The package **Sabre R** is available and comparable with the package **Sabre STATA**, but both packages do not include Gamma and nonparametric cloglog mixture models.

Finding the suitable software to estimate the parametric mixed proportional hazard models in Chapter 4 required a review for the packages that available in R and STATA software. Chart B.1 in Appendix B illustrates the available packages in both R and STATA that we examined empirically to estimate the discrete time hazard model in Chapter 4. Our target, in addition of finding a software package that estimates the parametric (M)PH model easily, is to find a package that also produces prediction of the hazard rate that allow us to do more analysis into comparing the quality of the (M)PH predictions with the quality of the predictions of the kernel hazard estimators. So, GLAMM software in STATA is used to estimate the (M)PH models. The predicted coefficients of the (M)PH in STATA are imported to R software also in addition to the pointwise predicted discrete time hazard. This allows to improve the comparison between the kernel estimation technique and (M)PH method better.

1.8 Overview of the Thesis

The thesis is written in individual chapters basis and organised as follows:

Chapter 2 compares the estimates of the kernel conditional density function estimator in the MDT estimation framework with the multinomial logit MNL parametric model estimates. The example attempts to estimate a female labour force participation model using data form UK labour market survey in 2007. A restricted version of the female labour force participation model is discussed first in the chapter. This highlights the differences between the predicted conditional probabilities of the MNL parametric method and kernel nonparametric estimation method better, as will be illustrate in details in the chapter. Kernel smoothing tests are used to test the correct specification of the parametric multinomial logit model, and to test the equality of the kernel unconditional density estimates of the variables that included in the model. The conclusions of Chapter 2 are interesting to both nonparametric estimation methods and labour economics. Where the chapter presents a deep comparison between the conditional probabilities that estimated by the MNL method and the kernel nonparametric method. The kernel nonparametric estimation technique is found capable more to interpret the heterogeneity in female labour force participation and deliver results that are better than the results of the parametric MNL estimation technique.

In Chapter 3, a new kernel hazard estimators for single state discrete time transition is developed. The new estimator is extended from the existing continuous time kernel hazard estimators. So, a review of the kernel methods to estimate the continuous time unconditional and conditional hazard rate is presented in the chapter. The chapter justify the selection of the type of discrete ordered kernel function that smooths the duration time variable in the kernel hazard estimator. Where the assumption underlying the development of the discrete ordered kernel function have to be considered in the estimator. The last sections in Chapter 3 presents a simulation study that invented in this research to compare between the discrete time kernel hazard estimator and the parametric discrete time proportional hazard estimator. The competing risks random right censoring is simulated in the study. The duration time variable and the right censoring time variable are simulated from the Weibull and the Uniform distributions respectively and a numerical approximation method is used to control for the proportion of the right censored observations. A grouping scheme of the time is used and the comparison of the models in the simulation study is allowed in the continuous time setting and the discrete time setting jointly. However, the simulation study is to compare between homogeneous duration model only, where no frailty is simulated in the DGP.

Chapter 4 includes an empirical study that compares between the results of the discrete time kernel hazard estimators that developed in Chapter 3 and the discrete time proportional hazard method. However, because the work uses a real data on duration of job vacancies in weeks, the mixed proportional hazard (MPH) models are included in the comparison to have models that control for the effect of the unobserved heterogeneity. There are repeated arguments in the literature of kernel nonparametric estimation methods that says that the nonparametric kernel estimators are less affected by the unobserved heterogeneity problem. So, the results of the kernel hazard estimator are compared with the results of homogeneous proportional hazard estimator and the mixed proportional hazard estimator in an attempt to examine this argument. The hazard rate is estimated for job vacancies in weeks using a data from Lancashire Careers Service (LCS) for the period from March 1988 to June 1992. The job search random matching model is used to compare between the estimation techniques, where the hazard of the job vacancies are estimated conditional on the number of job seekers and the number of job vacancies in the job market. It is found that the kernel hazard estimator produces results that consistent are with the random matching theory better than the parametric models. The work on chapters 3 and 4 refers to Cameron and Trivedi (2005) and Jenkins (2005) heavily in the definition of the teams that used in the discussion.

The last chapter, Chapter 5, includes a summary of the conclusions of the work in all the chapters in the thesis. Each chapter in the thesis has its own conclusions section that relates only to the work in that chapter. The final chapter presents an overview and an assessment for the performance of kernel method with fixed bandwidth and the mixed data types estimation framework in estimating DCF. The chapter presents useful suggestions for future research in this field in econometrics.

The Appendices are sorted as follows: Appendix A presents the key notation to the equations and the formulas. Appendix B is related to the discussion regarding the software packages that estimate mixed proportional hazard models in this chapter. Appendix C and D are related to Chapters 2 and 3 respectively. The results that thought to be too long to be presented inside those two chapters are presented in their relevant appendices. This include also basic component to the work that we think they may be too daunting for the reader if they are presented in the chapters, for example the mathematical proofs in Appendix D that are the fundamentals of the DGP design in Chapter 3.

Chapter 2

Nonparametric Conditional Density Analysis of Female Labour Supply Using UK Labour Force Survey for 2007

Chapter Abstract

This paper compares between kernel conditional density function estimation technique and multinomial logit MNL discrete choice parametric estimation technique. The model that is estimated is a female labour market participation decision model, using a cross section data from the labour force survey in the UK in 2007. Nonparametric smoothing specification tests for the discrete choice models, which are recently developed in econometrics, are used to compare between the two estimation techniques.

The kernel estimation method shows an outstanding performance compared with parametric MNL estimation method. This is true even when the MNL is improved by using semi-parametric specifications that relax the functional form of the continuous variables in the model. The smoothing specification tests show that the MNL models with different specification cannot be as good as the kernel method into capturing the heterogeneity in the propensity to work among the females in the sample. Attempting to improve the parametric models to make them be as powerful as the kernel estimators into capturing the heterogeneity may make them exhibit problems like multicollinearity.

The results of this research are interesting in labour economics. Nonparametric estimation method shows how female propensity to work is higher for females with no dependent children and high qualification levels. The propensity to work decreases with the increase in the number of dependent children, but for the females at age older than 40 years old the propensity is high irrespective to the number of dependent children, possibly due to the higher working experience for the females of that age. This last property in females propensity to work is captured by the kernel estimators only.

Key Words: Female labour force participation decision, nonparametric, discrete choice model, multinomial logit, binomial logit, kernel conditional density, mixed data types, maximum likelihood cross validation, delta method.

JEL Classification: C14, C21, C25, C41, C63, J10, J20, J24.

2.1 Introduction

Estimating models of female labour supply - in particular the decision whether or not to participate in the labour force - is usually analysed using discrete choice modelling techniques. In this literature, it is standard to explain the decision to participate using fertility variables (such as the number of dependent children or the number of children younger than 5), education, marital status and household status (for example, whether the female is the head of the household). In the keeping with this thesis, in this chapter we focus on nonparametric methods to estimate a female's participation decision; in particular, we use kernel conditional density with mixed data types estimation method and compare the results with the often-used discrete choice estimation technique, namely the multinomial logit. We use a sample of females of working age (16-59) from the UK Labour Force Survey in 2007, and estimate the conditional probabilities for whether the female is either employed, unemployed or economically inactive. We think that this is the first application of nonparametric kernel techniques for modelling female labour supply decisions using UK data.

The kernel conditional density estimator for mixed data types (MDT) is a newly developed nonparametric estimator that is considered as a major advancement in the nonparametric estimation in econometrics. This estimator is developed in Li and Racine (2003) and Racine et al. (2004). It is particularly attractive because, hitherto, the investigator was unable to specify conditioning variables of different "types", namely continuous covariates, discrete categorical unordered covaraites, and discrete ordered covariates. Where there were no suitable kernel functions for categorical variable in kernel estimators. For parametric estimation methods such as the multinomial logit, this issue is handled easily, but for kernel nonparametric estimation methods the problem of having different data types is "solved" in the mixed data framework by adopting a family of discrete unordered and discrete ordered kernel functions to smooth such variables.

The chapter is organised as follows; Section 3.2 presents a review of the literature that uses the discrete choice method in estimating female labour force participation models. Section 3.3 presents the econometric estimation method, and is divided into a presentation of the development of the parametric estimation technique, then a presentation of the kernel estimation technique and kernel smoothing tests. The sample is discussed in section 3.4, and the results of the models are discussed in section 3.5. The final section discusses the conclusions.

2.2 Modelling Female Labour Force Participation

The discrete choice modelling technique in econometrics was developed by Fishbein (1967), McFadden (1973), Manski and McFadden (1981), and Maddala (1983). An advanced discussion is found in econometrics textbooks like Wooldridge (2002), Greene

(2003), and Cameron and Trivedi (2005). Applications of discrete choice models in economics are extremely commonplace, including estimating models for consumer choice analysis, revealed and stated preferences analysis, transportation problems, willingness to pay, experimental economics and economics of happiness studies.

The **multinomial logit (MNL)** structure theoretically is the easiest structure among the discrete choice modelling structures. Where it is designed to estimate conditional probabilities for non-nested unordered alternatives like the labour supply choices of individuals, under the assumption that the unobserved utility has an extreme value distribution. Compared with other structures for non-nested unordered alternatives like **multinomial probit (MNP)** and **extreme value** discrete choice structures, the MNL is flexible, easy to estimate and adaptable in advanced applications like in the bootstrap method. The MNL structure is also directly expendable theoretically like when using Taylor expansion to its logit function to do more analysis.

There is a vast empirical literature that uses the multinomial discrete choice method to estimate the female labour force participation conditioning on fertility and education. Examples include Hill (1983), Lehrer (1992), Heineck (2004) and Molina et al. (2007). Papers that condition on the number of children are Shapiro and Mott (1979), Johnson (1980), Carliner (1981), Lehrer and Nerlove (1986), Lehrer (1992), Nakamura and Nakamura (1994), Xie (1997), Jacobsen et al. (1999) and Miller and Xiao (1999). The main finding is that fertility has a negative effect on a female's propensity to work; specifically an increase in the number of dependent children decreases females tendency to enter the labour force and then employed. The papers that discuss the effect of both number of children (fertility) and education level to female labour force participation in UK, using the multinomial, is found in the work of Lehrer (1992), Bingley and Walker (1997, 2001) and Bingley et al. (2005). In Sudan a female labour market participation model is estimated by Maglad and Consortium (1998). The education level shows an increasing effect on the tendency of women to be in the labour force and be in employment.

In addition to the above paper, there are number of papers that compare female labour force participation in different countries. Calhoun (1994) compares the results of multinomial logit method across chosen EU countries including the UK with the results in the US. Cross-country comparison studies are found also in Colombino and Tommaso (1996), Tommaso (1999) and Bratti (2003). The last work shows a comparison between Italy and a few European countries in the impact of fertility and education on female labour force participation.

Studies that compare within parametric discrete choice techniques, like between MNL and MLP methods, are less common. Di Tommaso and Weeks (2000) discuss the differences between the multinomial discrete choice models and binary choice models using female labour force participation data and both logit and probit structures. Their research focus on modelling the decisions of labour market participation conditioning on the child bearing decision. The research outcomes show that the multinomial models are better than the binary models in representing the data, especially if enough care is paid to avoid multi-dimensionality problem. They show that extending the model to a 'higher dimensions' model, which is a term that used in the paper to refer to the multinomial model with high number of alternatives, encounters the curse of dimensionality problem. So, the research recommends not to use dependent variable with large number of alternatives (higher dimension).

Comparisons between parametric discrete choice models and nonparametric or semiparametric models empirically in estimating the conditional probabilities of labour supply alternatives are available in Gerfin (1996), Fernández and Rodríquez-Poo (1997) and Goodwin and Holt (2002) among many others. Those papers use nonparametric regression estimator for a binary dependent variable that of female labour force participation decision. After the development of kernel conditional density with mixed variables types estimator a few comparison studies are available. For example Kumar (2006) uses data on female choice of work from the US and compares the output of a **binomial logit (BNL) model** and a binomial probit model estimates with the output of the kernel conditional density estimator. The work compares the number of correct predictions of each technique, and shows that the nonparametric kernel method estimates more correct probabilities.

Froelich (2006) develops a nonparametric kernel regression technique that uses a new method of locally weighting the data. The estimator is defined as the **local logit kernel regression**. The empirical example in the paper uses Swedish female labour labour force participation data to compares between the results of the new kernel regression local logit estimator and the BNL model results. The conclusion of Froelich (2006) is the same as that of Kumar (2006), which shows that kernel estimator produces a better ratio of correct predictions. However, Froelich (2006), additionally, claims that kernel nonparametric estimators are better than the parametric binary choice estimator on dealing with unobserved heterogeneity in the data. Both works however, focus on the binary dependent variable case only, the application of the kernel nonparametric technique to a labour supply model with a multinomial dependent variable has not come to our knowledge.

2.3 The Econometric Techniques

2.3.1 The Unordered Multinomial Logit Model

The MNL model is an easy and direct parametric model that used to estimate conditional probabilities of unordered categorical variables. Define $m+1$ mutually exclusive alternatives that are exhaustive and finite. The scalar Y_i^d denotes the alternative that is chosen by individual i. The superscript d is used to define that Y_i^d is a discrete variable.

Then for each individual the probability $p_{ik} = Pr[Y_i^d = k]$ denotes the probability that individual i chooses alternative k, where $k = 0, 1, ...m$. So, Y_i^d a sample draw from a multinomial density function $f(y^d) = \prod_{k=0}^{m} p_{ik}^{\delta_{ik}}$, where δ_{ik} is a dummy variable that equals 1 if alternative k is selected and zero otherwise.

Let the observed sample contains the values $\{Y_i^d, \mathbf{X}_i\}_{i=1}^n$. The row vector of the mixed independent variables $\mathbf{X}_i$ include q continuous variables X_{is}^c, $s = 1, 2, ..., q$, and p discrete variables $X_{ir}^d, r = 1, 2, ...p$. To match the notation in the MNL model with the kernel MDT conditional density estimator, in the next subsection, we will illustrate the estimators when the covariates are very simple set. Specifically, when the continuous covariates are not transformed to define a quadratic or a polynomial form and each discrete variable has only two possible outcomes. In this simplest case $\mathbf{X}_i$ is the same in the parametric estimator and the kernel MDT estimator. Additionally, $\mathbf{X}$ does not include a column of ones, so that $\boldsymbol{\beta}_k$ is a column vector of length $p + q$ that does not include an intercept.

The **random utility theory (RUT)** is used to construct the unordered multinomial discrete choice model. The RUT assumes that individuals seek utility maximisation by choosing appropriately from the basket of the alternatives. Let $U_{ki} = V_{ki} + \epsilon_{ki}$ define the utility of individual i that is obtained from alternative k. V_{ki} is defined as the captured part of the utility, the **representative utility**, that can be estimated in the model using $\mathbf{X}_i$ and the coefficient vector that is associated with alternative k, $\boldsymbol{\beta}_k$. ϵ_{ki} is the uncaptured part of the utility, or the part that captures the factors that are not included in V_{ki}. The probability for individual i to choose alternative k is given as follows:

$$
\begin{aligned}
p_{ik} &= \Pr\left[\epsilon_{ji} - \epsilon_{ki} < V_{ki} - V_{ji}; j \neq k\right] && (2.1)\\
&= \int_{\epsilon} I\left(\epsilon_{ji} - \epsilon_{ki} < V_{ki} - V_{ji}; j \neq k\right) f\left(\epsilon_i\right) d\epsilon_i, && (2.2)
\end{aligned}
$$

where $f\left(\epsilon_i\right)$ is the density function of ϵ_i.

The estimator of the conditional probability of choosing the k^{th} alternative by the i^{th} individual is found by specifying the distribution of ϵ_{ki}, and thus on the distribution of the difference of the unexplained part of the two alternatives $\epsilon_{ji} - \epsilon_{ki}$. By taking the integration to find the cumulative distribution of ϵ_i, the model is written as follows, Cameron and Trivedi (2005: page 496):

$$p_{ik} = \Pr\left[Y_i^d = k|\mathbf{X}_i; \boldsymbol{\beta}\right] = F_k\left(\mathbf{X}_i, \boldsymbol{\beta}\right), \quad k = 0, 1, 2, ..., m \quad i = 1, 2, ..., n, \quad (2.3)$$

$$F_k\left(\mathbf{X}_i, \boldsymbol{\beta}\right) \in [0, 1] \quad \text{and} \quad \sum_{k=0}^{m} F_k\left(\mathbf{X}_i, \boldsymbol{\beta}\right) = 1, \quad (2.4)$$

where $F\left(\cdot\right)$ is a link function that transforms the explained part of the utility into a probability. The parameters are estimated using the likelihood function:

$$\ell\left(Y_i^d, \mathbf{X}_i, \boldsymbol{\beta}\right) = \sum_{i=1}^{n} \sum_{k=0}^{m} \delta_{ik} \ln p_{ik}. \tag{2.5}$$

The link function must be identifiable and measurable for the independent variables and the parameters, and should be continuous in $\mathbf{X}$ and $\boldsymbol{\beta}$ values. The explanatory variables are not alternative specific and they are observed for all y^d values, while the parameter vector $\boldsymbol{\beta}_k$ is alternative specific and the model is normalised so that $\boldsymbol{\beta}_0 = 0$, where the alternative that $Y_i^d = 0$ is known as the **reference alternative**. The coefficient vector $\boldsymbol{\beta}_k$ is estimated by maximizing the likelihood function in Eq (2.5), which is available in most econometrics software.

The unordered multinomial logit MNL is defined when ϵ_i is assumed has an Extreme Value distribution with density function:

$$f(\epsilon_{ki}) = \exp\left(-e^{-\epsilon_{ki}}\right).e^{-\epsilon_{ki}}, \tag{2.6}$$

and distribution function

$$F(\epsilon_{ki}) = \exp\left(-e^{-\epsilon_{ki}}\right). \tag{2.7}$$

The multinomial logit MNL is easier than other specifications like the multinomial probit or generalised extreme value models and well established in the literature. The MNL link function is given by:

$$p_{ik} = \frac{\exp\left(\mathbf{X}_i \boldsymbol{\beta}_k\right)}{1 + \sum\limits_{l=1}^{m} \exp\left(\mathbf{X}_i \boldsymbol{\beta}_l\right)}.$$

Which is estimated as follows:

$$\widehat{p}_{ik} = F(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}) = \frac{\exp\left(\mathbf{X}_i \widehat{\boldsymbol{\beta}}_k\right)}{1 + \sum\limits_{l=1}^{m} \exp\left(\mathbf{X}_i \widehat{\boldsymbol{\beta}}_l\right)}. \tag{2.8}$$

The **maximum likelihood (ML)** coefficients are consistent and follow the normal distribution asymptotically as follows:

$$\widehat{\boldsymbol{\beta}} \overset{a}{\sim} N\left(\boldsymbol{\beta}, -E\left[\partial^2 \ell\left(y^d, \mathbf{X}, \boldsymbol{\beta}\right)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'\right]^{-1}\right). \tag{2.9}$$

The multinomial logit link function is twice continuously differentiable in $\mathbf{X}$ and $\boldsymbol{\beta} = [\mathbf{0}, \boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_m']$, McFadden (1984), Greene (2003) and Cameron and Trivedi (2005).

The changes in the conditional probabilities of the alternatives due to the change in the independent variables are known as the **marginal effects**. They are estimated from the partial derivatives of the link function with respect to the continuous independent variables, and by differencing between the probability at value 1 and the probability at value 0 for the dummy independent variables. When the marginal effect is estimated for some independent variable in the model, all other covariates are kept fixed at known values. The regular choice is to fix other covariates at the mean values, but different measures like the quartiles, the median, $\cdots$ etc, are also valid to use.

The marginal effects for continuous explanatory variables take the following form, Greene (2003, page 722):

$$\frac{\partial p_{ik}}{\partial \mathbf{X}_i} = p_{ik}\left[\boldsymbol{\beta}_k - \sum_{l=1}^{m} p_{il}\boldsymbol{\beta}_l\right], \tag{2.10}$$

which is estimated by $\widehat{\pi}_k$ with an asymptotic variance:

$$Asy.Var\left(\widehat{\pi}_{ik}\right) = \sum_{t=1}^{m}\sum_{l=1}^{m}\left(\frac{\partial \pi_{ik}}{\partial \boldsymbol{\beta}_t'}\right) Asy.Cov(\widehat{\boldsymbol{\beta}}_t, \widehat{\boldsymbol{\beta}}_l)\left(\frac{\partial \pi_{ik}}{\partial \boldsymbol{\beta}_l'}\right)'. \tag{2.11}$$

The **odds ratio**, or the **relative risk**, R, for alternative k rather than the reference alternative, is the probability of alternative k divided by the probability of the reference alternative,

$$R_{k0} = \frac{p_k}{p_0} = \exp(\mathbf{X}_i\boldsymbol{\beta}_k). \tag{2.12}$$

for the i^{th} individual. For any other two alternatives, say k and l, the relative risk is estimated by dividing the probabilities of the alternatives, which produces a formula for the difference between the alternative specific coefficient vector of k and l, $R_{kl} = \frac{p_k}{p_l} = \exp(\mathbf{X}_i\left(\boldsymbol{\beta}_k - \boldsymbol{\beta}_l\right))$.

The logarithm of the relative risk is practically useful in applied research since

$$\frac{\partial \log R_{k0}}{\partial \mathbf{X}} = \boldsymbol{\beta}_k,$$

and

$$\frac{\partial \log R_{kl}}{\partial \mathbf{X}} = \boldsymbol{\beta}_k - \boldsymbol{\beta}_l.$$

Then, a positive coefficient is interpreted as an increase in the probability of the alternative relative to the probability of the reference alternative when the independent variable increases, and the negative coefficient is interpreted as a decrease in the probability of the alternative relative to the reference alternative. This is in the dummy variables case is interpreted as an increase and a decrease in the relative probabilities from the individuals in the base-group for the positive and negative coefficients respectively. When the relative risk is computed for alternatives k and l the interpretation depends on the sign of the difference.

The **binomial logit BNL model** is used when the dependent variable is binary, has only two possible outcomes. This is a special case of the multinomial logit, where one of the outcomes is chosen as the reference alternative. The binomial logit model is simpler than the multinomial logit model. The interpretation of the coefficients is easier and direct more than the interpretation of the coefficient vector and the marginal effects of the multinomial logit.

The point-wise standard errors for the multinomial logit predicted probabilities are needed to compare the 95% confidence bands of $\widehat{p}_{ik}$with the corresponded 95% bootstrap confidence bands of the kernel nonparametric conditional density predicted probabilities. The Delta method is used to expand the logit link function, the point-wise variances are found as follows, for the reference alternative:

$$Var(\widehat{p}_{i0}) = \widehat{p}_{i0}^{2} \left\{ \left(R_i^{(0)} \otimes \mathbf{X}_i \right)^T \left[Var(\widehat{\boldsymbol{\beta}}) \right] \left(R_i^{(0)} \otimes \mathbf{X}_i \right) \right\}, \tag{2.13}$$

for other alternatives:

$$Var(\widehat{p}_{ik}) = \widehat{p}_{ik}^{2} \left\{ \left(R_i^{(k)} \otimes \mathbf{X}_i \right)^T \left[Var(\widehat{\boldsymbol{\beta}}) \right] \left(R_i^{(k)} \otimes \mathbf{X}_i \right) \right\}. \tag{2.14}$$

where:

$$R_i^{(0)} = \left[\widehat{p}_{1i}, \widehat{p}_{2i}, ..., \widehat{p}_{mi} \right]^T.$$

$R_i^{(j)}$ is $R_i^{(0)}$ with the probability of the alternative j replaced by its complement $(1 - \widehat{p}_{ik})$, and $\otimes$ denotes matrix Kronecker product.

2.3.2 Kernel Conditional Density Method for a Discrete Unordered Dependent Variable

The Development of the Kernel Conditional Density Estimator

The kernel smoothing method is suggested by Rosenblatt (1956) to estimate univariate density functions of continuous variables. The method is extended to estimate univariate probability mass functions for discrete variables by Aitchison and Aitken (1976) and improved in the work of Habbema et al. (1978), Titterington (1980), Wang and Van Ryzin (1981) and Aitken (1983), among many others, to different kinds of discrete variables. A clear distinction between discrete unordered and discrete ordered variables in the development of the estimators is followed. Kernel method to estimate multivariate density functions for mixed variables types attracts attention since the early development of kernel density estimation framework. Ahmad and Cerrito (1994) provides a method to estimate a bivariate density with one continuous variable and one discrete variable. Li and Racine (2003) and Racine et al. (2004) extend this estimator to the multivariate density with a number of discrete and continuous variables, which is the work that became highly distinguished in the kernel density estimation framework lately.

The development of the kernel estimator of conditional density functions started later than that of the unconditional kernel density estimators. Hyndman et al. (1996) for example, is one of the first contributors on this field, and the conditional density estimator that developed in his work uses the nonparametric kernel regression technique that was developed separately by Nadaraya (1965) and Watson (1964) in a conditional density context. Other work include Hall et al. (1999), Bashtannyk and Hyndman (2001), Hyndman and Yao (2002), Hansen (2004a,b), Hall et al. (2004), Fan et al. (2006) and Li and Racine (2008), with the asymptotic properties and the normality of the estimators discussed and established in Holmes et al. (2007) and Ezzahrioui and Ould-Said (2008) among many others.

The objective is to estimate the conditional density of y^d on $\mathbf{X}$. The kernel estimation approach does not require information about the utility function of the individuals. Instead, kernel estimator of the conditional density requires the estimators of the joint density and the marginal density functions. Where the conditional density is the ratio between the estimated joint density $\widehat{f}_{Y,X}(Y_i^d, \mathbf{X}_i)$ and the estimated marginal density $\widehat{f}_X(\mathbf{X}_i)$ as shown in Eq(2.20).

The joint density estimator at $Y_i^d = k$ and $\mathbf{X}_i$ is the average of the product of the univariate local weights that given to variables by the univariate kernel functions as follows:

$$\widehat{f}_{Y,X}(Y_i^d = k, \mathbf{X}_i) = n^{-1}\sum_{j=1}^{n} l\left(Y_j^d, k, \widehat{\gamma}_0\right) \prod_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{js}^c - X_{is}^c}{\widehat{h}_s}\right) \prod_{r=1}^{p} l\left(X_{ir}^d, X_{ir}^d, \widehat{\gamma}_r\right). \quad (2.15)$$

where $\widehat{\gamma}_0$ is the bandwidth of the discrete dependent variable, $\widehat{h}_s$ are the bandwidths of the univariate continuous variables, all presented in the vector $\widehat{\mathbf{h}}_{x^d} = [\widehat{h}_1, ..., \widehat{h}_q]$. $\widehat{\gamma}_r$ are the bandwidths of the and univariate discrete variables, presented in the vector $\widehat{\boldsymbol{\gamma}}_{x^d} = [\widehat{\gamma}_1, ..., \widehat{\gamma}_p]$. Then the vector of the bandwidths of the joint density function is $[\widehat{\gamma}_0, \widehat{\mathbf{h}}_{x^d}, \widehat{\gamma}_{x^d}]$. The bandwidth control the degree of smoothness of the function and are estimated jointly in the model using data driven method. The marginal density estimate on the other hand is the average of the product of the univariate local weights of the variables in $\mathbf{X}$ only as follows,

$$\widehat{f}_X(\mathbf{X}_i) = n^{-1}\sum_{j=1}^{n}\prod_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{js}^c - X_{is}^c}{\widehat{h}_s}\right) \prod_{r=1}^{p} l\left(X_{jr}^d, X_{ir}^d, \widehat{\gamma}_r\right), \quad (2.16)$$

the bandwidth vector of the marginal density function estimator is $\widehat{\mathbf{h}} = [\widehat{\mathbf{h}}_{x^c}, \widehat{\boldsymbol{\gamma}}_{x^d}]$, has length $p + q$ equals the length of the alternative specific coefficient $\widehat{\boldsymbol{\beta}}_k$, since we present the estimator for simple $\mathbf{X}$ as defined above. When the MNL model defines a quadratic or a polynomial form or one or more independent discrete variable has more that two possible outcomes, $\widehat{\mathbf{h}}$ and $\widehat{\boldsymbol{\beta}}_k$ will have different lengths. In the model that we estimate in this research the number of coefficients in $\widehat{\boldsymbol{\beta}}_k$ is larger than the number of the bandwidths of the marginal density $\widehat{\mathbf{h}}$ because we use a MNL with quadratic form and some discrete categorical variables have more that two possible outcomes.

The local weights are computed using the kernel function $w(\cdot)$ for the continuous variable that, in this research, is defined as a Second-Order Gaussian kernel function

$$w(z) = (2\pi)^{-\frac{1}{2}} \exp(-z^2/2). \tag{2.17}$$

$l(.,.,.)$ is a discrete variable kernel function that is defined differently for the discrete unordered variables and the discrete ordered variables. In the first case the unordered Aitchison and Aitken (1979) kernel function is used, which has the form:

$$l\left(X_i^d, x^d, \gamma\right) = \begin{cases} 1-\gamma & \text{if } X_i^d = x^d \\ \gamma/(c-1) & \text{if } X_i^d \neq x^d \end{cases}, \tag{2.18}$$

which is equivalent to;

$$l\left(X_i^d, x^d, \gamma\right) = [1-\gamma]^{\delta_i} \left[\frac{\gamma}{(c-1)}\right]^{1-\delta_i}, \quad \delta_i = 1\left(X_i^d = x^d\right).$$

For the ordered variables we use Wang and Van Ryzin (1981) kernel that takes the form:

$$l\left(X_i^d, x^d, \gamma\right) = \begin{cases} 1-\gamma & \text{if } X_i^d = x^d \\ \frac{1}{2}(1-\gamma)\gamma^{\left|X_i^d - x^d\right|} & \text{if } X_i^d \neq x^d \end{cases} \quad \gamma \in [0,1]. \tag{2.19}$$

See, Racine et al. (2004), Racine and Hayfield (2008) and Hayfield and Racine (2008, page 24).

So, the kernel conditional probability mass function estimate of $Y_i^d = k$, conditioning on $\mathbf{X}_i$, takes the following form, see Li and Racine (2007, chapter 5):

$$\begin{aligned} \widehat{f}_{Y|X}\left(Y_i^d = k \,|\mathbf{X}_i\right) &= \frac{\widehat{f}_{Y,X}(Y_i^d = k, \mathbf{X}_i)}{\widehat{f}_X(\mathbf{X}_i)}, \\ &= \frac{n^{-1}\sum\limits_{j=1}^{n} l\left(Y_i^d, k, \widehat{\gamma}_0\right) \prod\limits_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{js}^c - X_{is}^c}{\widehat{h}_s}\right) \prod\limits_{r=1}^{p} l\left(X_{ir}^d, X_{ir}^d, \widehat{\gamma}_r\right)}{n^{-1}\sum\limits_{j=1}^{n}\prod\limits_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{js}^c - X_{is}^c}{\widehat{h}_s}\right) \prod\limits_{r=1}^{p} l\left(X_{jr}^d, X_{ir}^d, \widehat{\gamma}_r\right)} \end{aligned} \tag{2.20}$$

where the bandwidths $\widehat{\gamma}_0$ and $\widehat{\mathbf{h}}$ are estimated combined together for the joint and marginal densities.

$\widehat{f}_{Y|X}\left(y^d \,|\mathbf{x}\right)$ estimator converges in distribution to a normal distribution as follows, Racine et al. (2004) and Li and Racine (2007, page 180):

$$\begin{aligned} &\left(n\widehat{h}\right)^{1/2} \left\{ \widehat{f}_{Y|X}\left(y^d \,|\mathbf{x}\right) - f_{Y|X}\left(y^d \,|\mathbf{x}\right) - B_1\left(\mathbf{x}, y^d\right)\widehat{h}^2 - \sum_{r=1}^{2} B_{2r}\left(\mathbf{x}, y^d\right)\widehat{\gamma}_r \right\} \\ &\xrightarrow{d} N\left(0, \left(\int w^2(z)\,dz\right) \frac{f_{Y,X}(y^d, \mathbf{x})}{f_X(\mathbf{x})}\right), \end{aligned} \tag{2.21}$$

where $B_1\left(\mathbf{x}, y^d\right)$ and $B_2\left(\mathbf{x}, y^d\right)$ are shown in Li and Racine (2007) Solutions Manual[1].

[1]Note that the formulas that are presented in Li and Racine (2007, Solutions Manual) illustrate the

The dependent variable is smoothed in the joint density function with the discrete unordered kernel function $l\left(Y_i^d, k, \widehat{\gamma}_0\right)$, and each observation Y_i^d in the sample is locally weighted based on whether it equals k or not, but without defining a reference value (alternative) in the kernel estimators. Then, in contrast to the MNL models, where a reference alternative in the dependent variable is needed, in the kernel nonparametric estimator the definition of the reference alternative is absolutely not necessary. The redefining of the reference alternative in the parametric MNL model leads to a reparameterisation of the coefficients of the MNL model while the estimated conditional probabilities and the estimated marginal effect coefficients remain the same. In the kernel estimator there is no such treatment, all the alternatives are equally smoothed in the model and matched with k with no distinction. Hence, it is easier to present the results of the parametric model in a form of predicted probabilities and marginal effect coefficients to make the comparison with the kernel method estimated results.

Bandwidth Selection Method

The bandwidths of the conditional density estimator are estimated using data-driven methods, like the **cross-validation method**, as suggested by Habbema, Hermans and Van den Broek (1974) and Duin (1976). Where, the mixing of the continuous and discrete variables in the model makes bandwidth approximation methods, like the Plug-in method or the Rule-of-Thumb method, not feasible. The advantages of the CV method in the conditional density estimation are shown in Loader (1999), Bashtannyk and Hyndman (2001), Li and Racine (2003), Fan and Yim (2004), Hall et al. (2004), Horne and Garton (2006) and Li and Racine (2007). In the models with mixed types variables, Li and Zhou (2005) demonstrates that the cross validation bandwidths have nicer convergence behaviour to the true smoothing parameters than the bandwidths that estimated by other methods, and show that the discrete variables in the model contribute in accelerating the convergence rate to the true bandwidths.

Hall et al. (2004) suggests two cross validation methods for the kernel conditional density estimator, the least squares cross validation (LSCV) and the maximum likelihood cross validation (MLCV). The two methods produce similar bandwidths when the data is well behaved, but the LSCV bandwidths have nicer asymptotic properties and are better recommended. The MLCV on the other hand, apply less numerical calculations and are easier to compute in models with many variables and a large sample size. However, the MLCV method may estimate bandwidths that over-smooth the density if the continuous variables in the model are drawn from fat tailed distributions, see Li and Racine (2007, page 161). Despite the fact that the good qualities of the LSCV bandwidths are more attractive in the model than the qualities of MLCV, we use the MLCV method to estimate

asymptotic properties of the estimator when both relevant and irrelevant regressors are included. In this research we show the same formula but when relevant regressors only are included. The reader should consider this difference when reading the formulas of $B_1\left(\mathbf{x}, y^d\right)$ and $B_2\left(\mathbf{x}, y^d\right)$.

the bandwidths because the LSCV bandwidths estimation is computationally very taught in our data.

The MLCV maximises an objective function that based on the Kullback-Leibler distance between the kernel density estimator and the true density. The leave-one-out (LOO) method is needed in the objective function. The objective function of the MLCV method is given in Li and Racine (2007, page 161) as:

$$L = \sum_{i=1}^{n} \log \widehat{f}_{-i}(Y_i \,|\mathbf{X}_i) = \tag{2.22}$$

$$\sum_{i=1}^{n} \log \left\{ \frac{\sum\limits_{j\neq i, j=1}^{n} l\left(Y_j^d, Y_i^d, \widehat{\gamma}_0\right) \prod\limits_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{js}^c - X_{is}^c}{\widehat{h}_s}\right) \prod\limits_{r=1}^{p} l\left(X_{jr}^d, X_{ir}^d, \widehat{\gamma}_r\right)}{\left[\sum\limits_{j\neq i, j=1}^{n} \prod\limits_{s=1}^{q} \widehat{h}_s^{-1} w\left(\frac{X_{js}^c - X_{is}^c}{\widehat{h}_s}\right) \prod\limits_{r=1}^{p} l\left(X_{jr}^d, X_{ir}^d, \widehat{\gamma}_r\right)\right]} \right\} \tag{2.23}$$

Without the LOO method the CV method collapses and the estimation of the bandwidths are not attainable, see Li and Racine (2007, page 16). The LOO method uses each observation in the sample n times to compute the value of the objective function at each iteration. For each individual in the sample the conditional density $\widehat{f}(Y_i \,|\mathbf{X_i})$ is estimated using all the observations except the observation that related to the i^{th} individual, i.e by using $n-1$ observations in the sample. This estimates the conditional density with one observation left out each time, then the objective function is the sum of the logarithms of all the LOO estimated densities.

2.3.3 Nonparametric Kernel Tests

Discrete Choice Model Specification Tests

The kernel nonparametric smoothing technique is used to test the correct specification of many parametric models. The kernel smoothing tests is a broad framework that include families of tests that some of them apply the bootstrap method, a addition to the smoothing method, to calculate the test statistic and its empirical distribution. Kernel correct specification smoothing tests are available for regression function and continuous variable density function estimator, see Li and Racine (2007, Chapter 12), and they are widely considered in the literature and many tests of this type are now available in many software packages. Tests for the correct specification of the discrete choice model are not very common in the literature compared with the tests of correct specification of parametric regression functions. Where only the tests that developed by Fan et al. (2006) and its duplicate test that developed by Zheng (2008) appeared to us. Both tests extend the correct specification test of Zheng (2000) to the discrete choice model case. The bandwidths that are estimated for the kernel conditional density are used to calculate the test statistic. The bootstrap method is required to approximate the distribution of the test statistic under the null empirically.

The tests are one sided and the test statistic tends asymptotically to a standard normal distribution. The null hypothesis is that the parametric discrete choice model is correctly specified. The critical value is calculated using the parametric bootstrap method from the discrete choice model using the method of Andrews (1997). The bootstrap procedure is identical in the two tests. However, the variance of the distance measure in the test is different in Fan et al. (2006) to that in Zheng (2008), so that the standardised version of the test statistic is also different. Zheng (2008) claims that the test is robust and "has high power in all directions", but Fan et al. (2006) on the other hand argue that the test has the power to remove the irrelevant variables in the models and "enjoys substantial power gains in finite-sample application". There seems to be no application of these tests in empirical research.

The null and alternative hypotheses are, see Fan et al. (2006, page 590):

$$H_0 : Pr\left[f(Y_i^d\,|\mathbf{X}_i) = p\left(Y_i^d\,|\mathbf{X}_i\,,\boldsymbol{\beta}\right)\right] = 1,$$
$$H_1 : Pr\left[f(Y_i^d\,|\mathbf{X}_i) = p\left(Y_i^d\,|\mathbf{X}_i\,,\boldsymbol{\beta}\right)\right] < 1,$$

The tests are likelihood ratio type tests that are constructed from the kullback-leibler distance, the statistic is computed as follows:

$$U_n^* \quad = \quad \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j\neq i,j=1}^{n}\mathbf{W}\left(\mathbf{X}_i,\mathbf{X}_j,\widehat{\mathbf{h}}\right)\Upsilon_{ij}^*, \tag{2.24}$$

where $\mathbf{W}\left(\mathbf{X}_i,\mathbf{X}_j,\widehat{\mathbf{h}}\right)$ is the product of the kernel functions of the independent variables in the model:

$$\mathbf{W}\left(\mathbf{X}_i,\mathbf{X}_j,\widehat{\mathbf{h}}\right) = \prod_{s=1}^{q}\widehat{h}_s^{-1}w\left(\frac{X_{js}^c - X_{is}^c}{\widehat{h}_s}\right)\prod_{r=1}^{p}l\left(X_{jr}^d, X_{ir}^d, \widehat{\gamma}_r\right), \tag{2.25}$$

and

$$\Upsilon_{ij}^* = \frac{I\left(Y_i^{d^*} = Y_j^{d^*}\right) - p\left(Y_i^{d^*}\,|\mathbf{X}_j\,,\widehat{\boldsymbol{\beta}}^*\right)}{p\left(Y_j^{d^*}\,|\mathbf{X}_i\,,\widehat{\boldsymbol{\beta}}^*\right)}, \tag{2.26}$$

is an indicator function that depends on the discrete dependent variable and the predicted probabilities of the parametric model. The superscript $(*)$ indicates that the dependent variable $Y_i^{d^*}$ is simulated from the predicted probabilities of the MNL, and the coefficients $\widehat{\boldsymbol{\beta}^*}$ are estimated from the bootstrap samples. The estimated variance of the statistics are different, which makes the standardised value of the tests different. In Fan Test the variance of the test statistic is

$$\widehat{\sigma}_{Fan}^{2*} = \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j\neq i,j=1}^{n}\left(\mathbf{W}\left(\mathbf{X}_i,\mathbf{X}_j,\widehat{\mathbf{h}}\right)\Upsilon_{ij}^*\right)^2, \tag{2.27}$$

but the variance of Zheng Test statistic is

$$\widehat{\sigma}^2_{Zheng} = \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j\neq i, j=1}^{n} \mathbf{W}\left(\mathbf{X}_i, \mathbf{X}_j, \widehat{\mathbf{h}}\right)^2. \tag{2.28}$$

The test statistics from the bootstrap samples are

$$\kappa^*_{Fan} = n(nh_1\cdots h_r)^{1/2} U^*_n / \widehat{\sigma}^*_{Fan}, \tag{2.29}$$

in Fan test, and

$$\kappa^*_{Zheng} = n(nh_1\cdots h_r)^{1/2} U^*_n / \widehat{\sigma}_{Zheng}, \tag{2.30}$$

in Zheng test.

The bootstrap method of Andrews (1997) is applied in this research using the following steps:

1 Use the cumulative conditional distribution of the dependent variable that is estimated by the MNL model $F(Y^d_i = k|\mathbf{X}_i, \widehat{\boldsymbol{\beta}}) = \sum\limits_{l=0}^{k} \widehat{p}_{il}$ to simulate the dependent variable Y^{d*}_i. The covariates are not simulated, thus the bootstrap sample contain $\{Y^{d*}_i, X_i\}^n_{i=1}$. Y^{d*}_i is simulated as follows: a continuous $Uniform\,[0,1]$ random variable U^c_i is simulate, then Y^{d*}_i is set equals 0 if $U^c_i \leqslant F(Y^d_i = 0|\mathbf{X}_i, \widehat{\boldsymbol{\beta}})$ (in other words when $U^c_i \leqslant \widehat{p}_{i0}$ for the reference alternative in the MNL model). For other alternatives, Y^{d*}_i is set equals k if $F(k-1\,|\mathbf{X}_i\,, \widehat{\boldsymbol{\beta}}) < U^c_i \leqslant F(k\,|\mathbf{X}_i\,, \widehat{\boldsymbol{\beta}})$ for $k = 1, 2, ..., m$.

2 Use the bootstrap sample to estimate the bootstrap estimated coefficient vector, $\widehat{\boldsymbol{\beta}}^*$.

3 Use the bootstrap sample parametric model estimated coefficients, $\widehat{\boldsymbol{\beta}}^*$, and the cross-validation bandwidths vector that are estimated from the sample, $\widehat{\gamma}_0$ and $\widehat{\mathbf{h}}$, to compute the test statistics in Eq (2.24) and the variance in Eq (2.27). Where the variance of Zheng test in Eq (2.28) is computed from the original sample and not form the bootstrap samples.

4 Repeat steps (1) to (3) B times, in each bootstrap sample compute $\kappa^*_{(\cdot)}$, where $(\cdot)$ denotes Fan or Zheng test. The empirical distribution of the B bootstrap statistics $\{\kappa^*_{(\cdot)}\}^B_{b=1}$ is then used to approximate the distribution under the null.

5 Reject the H_0 at significant level α if the test statistic from the sample exceeds the α^{th} percentile of the empirical bootstrap distribution.

Zheng (2008) and Fan et al. (2006) show that, $(nh_1\cdots h_r)^{1/2} U^*_n / \widehat{\sigma}^*_{Fan} \xrightarrow{d} N\,(0,1)$, and $(nh_1\cdots h_r)^{1/2} U^*_n\ / \widehat{\sigma}_{Zhg} \xrightarrow{d} N\,(0,1)$ in Fan and Zheng tests respectively. The MLCV bandwidths that are estimated for the kernel conditional density are used to smooth the observations in the test statistic and the bootstrap samples.

In this research we apply the discrete choice specification tests after a simulation study that replicates some of the simulations of Zheng (2008) only but both tests are examined. In other words, we replicate some of Zheng (2008) simulations but adds Fan et al. (2006) test to our work. R software is used, the results are shown in Appendix C. The replication of all the results in the two papers is time consuming and not one of the objectives of this research, so we focus on replicating part of the results and reducing the number of Monte Carlo samples. Our simulation shows that Fan et al. (2006) and Zheng (2008) tests outcomes are very close. The size of the tests is estimated correctly in the replicated simulation, and close to the size of the test that shown in Zheng (2008). The power, on the other hand, is improved in our replicated simulation study, the power of Zheng (2008) test is higher than that shown in the paper. This may be due to the difference in the software package that used in our replication. This shows that our code is correctly written and valid to be used in the research.

Unconditional Densities Equality Test

The test of equality of the unconditional densities using kernel method for mixed data types is suggested by Li et al. (2009) and is available in R software in the NP package of Hayfield and Racine (2008). The null hypothesis of the test is that the two densities are equal, $H_0 : f(x) = g(x)$, is applied for the unconditional densities of discrete, continuous, univariate and mixed types multivariate densities. The kernel densities $\widehat{f}(\mathbf{x})$ and $\widehat{g}(\mathbf{x})$ are densities estimated in two different samples of size n_1 and n_2 respectively.

Denote the mixed variables type by $\{\mathbf{X}_i\}_{i=1}^{n_1}$ in the first sample and $\{\mathbf{Z}_i\}_{i=1}^{n_2}$ in the second sample. The test statistic is constructed based on the integrated squared density difference, Li et al. (2009, Eq 2.6), and has the following formula,

$$\begin{aligned} I_n &= \frac{1}{n_1\left(n_1-1\right)}\sum_{i=1}^{n_1}\sum_{j\neq 1}^{n_1} K_{\mathbf{h}}(X_i, X_j) + \frac{1}{n_2\left(n_2-1\right)}\sum_{i=1}^{n_2}\sum_{j\neq 1}^{n_2} K_{\mathbf{h}}(Z_i, Z_j) \\ &\quad - \frac{1}{n_1 n_2}\left(\sum_{i=1}^{n_1}\sum_{j\neq 1}^{n_2} K_{\mathbf{h}}(X_i, Z_j) + \sum_{i=1}^{n_2}\sum_{j\neq 1}^{n_1} K_{\mathbf{h}}(X_j, Z_i)\right), \end{aligned} \tag{2.31}$$

where $K_{\mathbf{h}}(X_i, Z_j)$ is a kernel product term $\prod_{s=1}^{q} h_s^{-1} w\left(\frac{X_{is}^c - Z_{js}^c}{h_s}\right) \prod_{r=1}^{p} l\left(X_{ir}^d, Z_{jr}^d, \gamma_r\right)$ and $K_{\mathbf{h}}(X_j, Z_i)$ switching the subscript, the vector of the bandwidths $\mathbf{h}$ is the bandwidth of the pooled sample $n_1 + n_2$. The variance of the test statistic is

$$\begin{aligned}\sigma_n^2 &= (2n_1n_2h_1, \cdots, h_r)\left[\frac{1}{n_1^2\,(n_1-1)^2}\sum_{i=1}^{n_1}\sum_{j\neq 1}^{n_1}K_{\mathbf{h}}^2(X_i, X_j)\right.\\ &\quad + \frac{1}{n_2^2\,(n_2-1)^2}\sum_{i=1}^{n_2}\sum_{j\neq 1}^{n_2}K_{\mathbf{h}}^2(Z_i, Z_j)\\ &\quad \left. - \frac{1}{n_1^2n_2^2}\left(\sum_{i=1}^{n_1}\sum_{j\neq 1}^{n_2}K_{\mathbf{h}}^2(X_i, Z_j) + \sum_{i=1}^{n_2}\sum_{j\neq 1}^{n_1}K_{\mathbf{h}}^2(X_j, Z_i)\right)\right],\end{aligned} \tag{2.32}$$

$\tau_n = (2n_1n_2h_1, \cdots, h_r)^{1/2} I_n^*/\widehat{\sigma}_n^* \xrightarrow{d} N\,(0, 1)$

The test statistics converges to a standard normal distribution as illustrated in Theorem 2.1 in the Li et al. (2009). n_1 and n_2 are used in Li et al. (2009) to develop the bootstrap procedure to approximate the test statistics under the null hypothesis. The pooled sample of n_1 and n_2 is used to independently draw two samples of size n_1 and n_2 respectively and then compute the test statistics.

The cross validation bandwidths are used in computing the test statistic and in the bootstrap distribution $\{\widehat{\tau}^*\}_{b=1}^{B}$, the $(1-\alpha)^{th}$ quantile is compared with the sample statistics $\widehat{\tau}$. As this test is provided in R software no replication to the simulations in Li et al. (2009) is needed.

2.4 The Data

In the empirical analysis, we analyse a sample of females of working age (16-59) drawn from the first quarter 2007 of the UK Labour Force Panel Survey. The sample, will be denoted as the **females in the working age sample**, comprises 35311 females. The dependent variable, Y^d, is the working status of the females. There are three working states that are defined: the **employed state**; for the individuals who are employed; the **unemployed state**, comprises the individuals who are seeking jobs; and the **economically inactive state**. The last group is the group are females who are said to be out of the labour force, and, by definition, are neither working nor seeking a job. The unemployed state includes all the individuals without a job but thinking or seeking to start work, continuously, for at least four weeks. They are individuals currently not working but able to move to the employed state in at least a couple of weeks period.

The independent variables that are included in our research are, mostly, some of the common covariates that are used the estimated models of female labour force participation decision in the literature that presented in Section (2.2). The covariates are the basic demographic and educational characteristic of the females, and those are directly related to female labour supply. There are 5 mixed independent variables: a continuous variable, Age, which is denoted by X_{age}^c; the Number of Dependent Children, X_{nc}^d, that defined as a categorical variable with 4 categories, no children, one child, two children and three or more children; the Education Level, X^d_{ed}, with 3 ordered categories defined as, the **High Qualification** group, including the females with university degree or higher qualifications, the **Medium Qualifications** group, the group that includes all the females with either a GCE a level or GCSE grades a-c or equivalent qualifications, and last the **Low Qualification** or the **GCE/GCSE** group, which is the group of all females with qualifications that lower than the women in the two first groups in addition to the individuals with no qualifications. Then a variable of whether the female in the sample is the head of the household, X^d_{hoh}, which defined as a categorical variable. Finally a categorical variable that shows whether the female has an under 5 years old child (or children), X^d_{u5c}.

We estimate a model using a subset that includes only: age, number of dependent children and education level (X^c_{age}, X^d_{nc}, and X^d_{ed}), which is called **the restricted model**. The model with all the covariates is called **the unrestricted model**. The results of the restricted model is presented in the chapter but not used in the kernel smoothing tests.

The sample is divided into smaller sub-samples using the economic activity dependent variable and the education level independent variable. First, all the females in the labour force are separated in one sub-sample, by dropping the group of the inactive females in the original sample. This generates a sample of size 26007 female, denoted by the **females in the labour force sample**. The dependent variable in the labour force sample is U^d that has only two categories: whether the individual is employed or unemployed. Each of the females in the working age sample and the females in the labour force sample is divided into three smaller sub-samples using the highest qualification variable, where each qualification level is used to generate a sub samples. The newly generated 6 sub-samples are as follows: females in the working age and high qualifications, females in the working age and medium qualifications, females in the working age and low qualifications, females in the labour force and high qualifications, females in the labour force and medium qualifications, females in the labour force and low qualifications. With the original sample the total number of sample and sub-samples that used in the research is 8. Accordingly, the dependent variable is multinomial in the females in the working age sample/sub-samples, and binomial in the samples/sub-samples of the females in the labour force.

The objective of the education level sub-samples is to overcome the computation cost of the kernel smoothing tests, since the empirical distribution of the tests is difficult to estimate in the **samples of all females**, which are the samples that include the females with all qualification levels, due to the tough calculations that are involved. The predicted probabilities in the pooled sample are matched with the predicted probabilities in the sub-samples and they are very similar.

The comparison includes two stages. The first stage compares the restricted models. Which are the models given by the estimator in Eq (2.20) when $q = 1$ and $p = 2$. The restricted models allow to examine the predicted probabilities in a small number of cells, since there are only 12 cells that are created by the cross categories of number of children and education level variables. This makes the number of observations in each cell sufficiently large to estimate the pointwise standard errors. Next is the comparison of the unrestricted models, in Eq (2.20) those are the models with $q = 1$ and $p = 4$. The number of cells in the unrestricted models is 48, generated from the cross categories of 4 discrete variables. The graphs illustrate the differences in the predicted probabilities of the estimation techniques in the restricted models better. While the smoothing specification tests illustrate the differences in the unrestricted models better.

2.4.1 The Summary Statistics and the Restricted Model Results

The descriptive statistics for the variables in the model are shown in Table 2.1. The summary statistics of the dependent variable shows that the majority of females are in the employed state, for the sample of females in the working age 16-59 the percentages are 69.9% in the employed state and 26.3% in the inactive state. The percentages of employed females increase as the qualification levels increases. In the sub-sample of females in the working age and high qualifications 85.4% of females are in the employed state, for those in the labour force 97.5% are in the employed state. This compares with the percentages 50.5% and 90.5% respectively in the sub-sample of females with low qualifications. The number of dependent children is higher for females with low qualifications, the proportion of mothers with dependent child/children under 5 years old is slightly different among the sub-samples but considerably different for the high and medium qualifications labour force sub-samples. The proportion of females who are heads of households is also similar in the sub-samples.

The statistics in Table 2.1 show differences in average age between females in qualification level groups, the average age of females with low qualifications is 6 years higher than the average age of females with medium qualification. Females with high qualifications have higher average age than females with medium qualifications by about 4 years. There is 4.5 percentage points drop in the proportion of females with low qualifications and three or more dependent children in the labour force sample.

The summary statistics in Table 2.1 show well behaved variables in the sample. The data is suitable to be used for comparison between parametric and nonparametric estimation techniques.

Restricted Parametric Models

Two specifications of the MNL and the BNL discrete choice models are estimated using a logit link function: one with a quadratic form in age and one with semiparametric form in age by using age dummy variables. The second model relaxes the restrictive functional form of age by using the dummy variables. In all logit models, the dummies of the categorical variables, the number of dependent children and the education level

Table 2.1: Summary statistics.

	Sample	All females	Highest Qualification sub-samples: High Qual.	GCE/GCSE	Low Qual
Panel A: Females in working age 16-59					
Sample size		35311	10125	16181	9005
Y^d	employed	0.699	0.854	0.709	0.505
	unemployed	0.038	0.022	0.039	0.053
	Inactive	0.263	0.124	0.252	0.442
X^c_{age}	Mean	38.23	39.52	35.63	41.44
X^c_{nc}	0 child	0.492	0.550	0.433	0.532
	1 child	0.220	0.192	0.249	0.201
	2 children	0.199	0.190	0.224	0.164
	3 or more	0.089	0.068	0.094	0.103
X^d_{hoh}		0.255	0.250	0.235	0.297
X^d_{u5c}		0.168	0.177	0.172	0.154
X^d_{ed}	Degree or Higher	0.287	-	-	-
	GCE/GCSE	0.458	-	-	-
	Low Qual.	0.255	-	-	-
Panel B: Females in the labour force					
Sample size		26007	8869	12111	5027
U^d	employed	0.949	0.975	0.948	0.905
	unemployed	0.051	0.025	0.052	0.095
X^c_{age}	Mean	38.66	39.30	36.60	42.50
X^c_{nc}	0 child	0.529	0.564	0.473	0.601
	1 child	0.219	0.194	0.244	0.202
	2 children	0.189	0.183	0.213	0.142
	3 or more	0.063	0.059	0.070	0.055
X^d_{hoh}		0.242	0.253	0.222	0.270
X^d_{u5c}		0.135	0.156	0.137	0.091
X^d_{ed}	Degree or Higher	0.341	-	-	-
	GCE/GCSE	0.466	-	-	-
	Low Qual	0.193	-	-	-

are included. The base group is defined as the group of females with no children and low education level. The estimated coefficients of the logit models are little useful in the comparison, so the marginal effects coefficients only are presented. The marginal effect coefficients are estimated at the mean of all and the sum of all covariate in the MNL logit model in all alternatives equals 0. The reference alternative is the employed state. For the models with semi-parametric form in age the marginal effect of age dummy variables are not reported for brevity.

Table 2.2 shows that the quadratic form is accepted in all samples, which indicates a concave shape on age for the conditional probability of the employed state and a convex shape on the inactive state. Negative coefficients show decreasing conditional probability of the alternative, like the coefficients of the number of children dummies on the probability of being in the employed state and the coefficients of education level in the conditional probability of being unemployed and being inactive. Positive marginal effects, in contrast, show an increasing effect on the conditional probability of the alternative, and are estimated for the education level in the employed state and number of children in the inactive state. In the MNL models, the majority of females in the sample are either employed in the labour force or inactive, only a handful number of females are in the unemployed state and, so the value of marginal effects of this alternative are close to zero.

Likelihood Ratio test statistics (LR test) at the bottom of Table 2.2 show that the parametric model with semi-parametric form in age is accepted against the model with the quadratic form, where the value of the test exceeds 204 in MNL model and 56 in the BNL model with high significance. The estimated marginal effect in the models with semi-parametric form in age are close to the marginal effects that are estimated to the models with the quadratic form. The value of the Log Likelihood shows a substantial improve in the models that relax the restrictive quadratic functional form on age.

The marginal effects show that females with three or more children have lower probability of being employed by about 36 percentage points than the females with no children. This is about three times the decrease in the probability for the females with one child, where their probability of being employed drops by about 10 percentage points only. Females with GCE/GCSE are 20 percentage points more likely to be employed than females with low qualifications, the base group. Females with high qualification on the other hand, are about 27 percentage points more likely to be employed than the low qualification females. This means that the difference between the probability of employed state among females with high qualification and females with GCE/GCSE qualification is only 7 percentage points. The table shows that the drop in the probability of being employed almost equals the increase in the probability of being inactive. The marginal effect of the number of dependent children in the unemployed state are almost zero. This magnitude appears in the model because the majority of the females with dependent children in the sample are inactive, and very small number of females are employed and having dependent children.

Table 2.2: Marginal effect for the restricted MNL and the restricted BNL models[1].

Model		MNL			**BNL**
Alternative		Employed	Unemployed	Inactive	Unemployed[4]
Panel A: Quadratic form in X_{age}^c					
$p\left((\cdot)^d \left\|\overline{\mathbf{X}}\right.\right)^2$		0.728	0.034	0.238	0.032
Age($= X_{age}^c$)		4.527 (0.140)	-0.366 (0.049)	-4.161 (0.130)	-0.228 (0.045)
Age Squared		-5.691 (0.187)	0.294 (0.067)	5.397 (0.174)	0.133 (0.063)
Number of Children[3] X_{nc}^d:					
	1 child	-0.105 (0.007)	0.003 (0.002)	0.103 (0.007)	0.002 (0.002)
	2 children	-0.173 (0.008)	-0.004 (0.002)	0.177 (0.008)	-0.005 (0.002)
	3 or more	-0.358 (0.011)	-0.001 (0.003)	0.359 (0.011)	-0.002 (0.003)
Education Level[3] X_{ed}^d:					
	gcs/gcse	0.201 (0.006)	-0.023 (0.002)	-0.178 (0.005)	-0.017 (0.002)
	high qual.	0.270 (0.005)	-0.029 (0.002)	-0.241 (0.005)	-0.025 (0.002)
Log likelihood		-22897.312			-5407.5366
n		35311			26007
Pseudo R^2		0.1058			0.0445
Wald stat.		4638.25			467.36
Panel B: Year dummies of X_{age}^c					
$p\left((\cdot)^d \left\|\overline{\mathbf{X}}\right.\right)^2$		0.730	0.033	0.237	0.038
Age dummies		the marginal effects of the age dummies are not shown			
Number of Children[2] X_{nc}^d:					
	1 child	-0.114 (0.008)	-0.0005 (0.002)	0.114 (0.008)	0.008 (0.003)
	2 children	-0.181 (0.009)	-0.007 (0.002)	0.188 (0.009)	0.003 (0.004)
	3 or more	-0.364 (0.011)	-0.004 (0.003)	0.368 (0.011)	0.024 (0.006)
Education Level[3] X_{ed}^d:					
	ges/gcse	0.200 (0.006)	-0.024 (0.002)	-0.176 (.005)	-0.037 (0.003)
	high qual.	0.270 (0.005)	-0.028 (0.002)	-0.241 (.005)	-0.045 (0.002)
Log likelihood		-22795.229			-4770.8475
n		35311			26007
Pseudo R^3		0.1098			0.0902
Wald stat.		4816.35			930.49

[1] Restricted model are the models that predict female labour force participation on age, number of dependent children and education level only.
[2] $p\left((\cdot)^d \left\|\overline{\mathbf{X}}\right.\right)$ indicates $p\left(y^d \left\|\overline{\mathbf{X}}\right.\right)$ in the MNL model and $p\left(u^d \left\|\overline{\mathbf{X}}\right.\right)$ in the BNL model.
[3] The basegroups are: for the Number of Children, the no children group; for the Education level, the Low Qualification group.
[4] The marginal effects of the BNL are for the probability of unemployment for the females in the labour force.

Restricted Kernel Nonparametric Models

The MLCV optimum bandwidths of the nonparametric kernel conditional density estimate are reported in Table 2.3. The values of the bandwidths suggest a nice kernel conditional density estimate for each of the multinomial dependent variable and the binomial dependent variable. Where all the bandwidths are sufficiently higher than zero, which removes the suspicion of a kernel density under-smoothing problem. The bandwidths show also that all the variables are relevant in the model, where the CV method estimates high bandwidth values for the irrelevant variables, and bandwidths that close to 1 for the irrelevant discrete variables. Fortunately the estimated bandwidths are not showing such a magnitude for any of the variables in the model.

The way to summary of the output of the kernel estimates is completely different than the way to summary the output of the parametric models. The bandwidths are not informative about the change in the conditional probabilities of the alternatives as the marginal effects coefficients in the parametric model. Compare the predictions of the parametric logit model with the prediction the kernel conditional density graphical presentation is used, because it is easy and effective approach. The conditional probabilities of the parametric models show a concave shape with age in the employed state and a convex shape in the inactive state. The change in the functional form of age in the model changes those shapes directly, so the probabilities are plotted with age in all figures to illustrate the effect of changing the specification of the functional form of the continuous variable in the parametric models better. For the categorical variables the parametric models predict the effect as a shift in the probabilities at all ages equally. So, choosing age variable to plot the probabilities makes the differences emerge clearly if the kernel estimator does not predict the effect of the categorical variables equally in all ages.

Table 2.3: Bandwidths of the restricted nonparametric kernel conditional density estimate of Y^d and U^d models in the pooled sample[1].

conditional pdf	Multinomial[2] conditional function	Binomial[2] conditional function
Bandwidth		
Y^d	0.001990	0.001134
Age X^c_{age}	1.647721	2.008208
Number of Children X^d_{nc}	0.081346	0.246962
Eduction level X^d_{ed}	0.020117	0.017490
CV. Optimum Value	0.640072	0.183654
n	35311	26007
CV iterations	-22398.87	-84241.91
log likelihood	3	1

[1] The bandwidth are estimated using the MLCV method.

[2] The types of the kernel functions are: Second Order Gaussian kernel for the continuous variables, Aitchison and Aitken (1979) kernel function for the unordered discrete variables and Geometrical kernel function for the ordered discrete variables.

Figures 2.1 and 2.2 show the plots of the predicted probabilities for the employed alternative and the inactive alternative respectively. The panels in the figures are, from top to bottom: the sample proportion of females in the economic activity state, the predicted conditional probabilities of the MNL with quadratic form in age, the predicted conditional probabilities of the MNL with age dummies, and the predicted conditional probabilities of the nonparametric kernel conditional density estimate. The predicted probabilities of the basegroup in each model is the red dotted line in the low qualifications column. The difference between the conditional probabilities that are predicted by the MNL model and the kernel conditional density estimator highly distinguishable.

The quadratic form MNL model makes the predicted probabilities extremely different from the probabilities that implied by the sample proportions.

The sample proportions show that females with different number of dependent children have higher propensity to work in all age groups than other females. Females below 40 years old are more likely to be economically inactive if they have dependent children and the proportion of being in the employed state drop substantially at these young working ages. For the group of females above 40 the differences in employment propensity decrease due to the increase in the propensity for the females with children, up to a level that make them equivalent to the propensity of females with no children.

Relaxing the quadratic functional form in age improves the parametric model, the semi-parametric form in Table 2.2 shows better Log Likelihood value and the LR test result shows that the semi-parametric form is accepted. The figures show that the predicted probabilities of the parametric models are improved when the semi-parametric form is used. The only improvement, however, is that the model becomes able to capture the diversity in the conditional probability with age and does not enforce a restrictive concave/convex shape in the model. Then the semi-parametric model does not imply a certain age at which the propensity to work for all females reach the peak. The parametric models enforce certain shapes for conditional probabilities of the females with different qualification. This contradicts with the behaviour that shown by the sample proportions. The conditional probabilities that predicted by the parametric method are very restrictive, and the parametric models failed to capture that females after age 40 have increasing propensity for work even with dependent children.

The kernel estimator, as shown in the plot, explains the heterogeneity in the model better than the parametric logit model. The conditional probabilities that are estimated by the kernel conditional density estimator show predicted conditional probabilities that are consistent with those implied by the sample proportions. The predicts of the kernel estimator is sensible in explaining the relationships between the variables in the model and the propensity of females to work. The advantage of the kernel estimate is that it shows this quality in the predictions without assuming any functional form in the model, and without using dummy variables or interaction terms. The number of the regressors in the kernel conditional density estimator is smaller than the number of covariates in the logit models, even though they specify the same set of independent variables. This shows that kernel estimation technique is more powerful than the parametric method.

The plots in Figures 2.3 and 2.4 compare between the precision of the predicted probabilities of the parametric model and the nonparametric kernel estimates in chosen cells. The 95% confidence bands of the probabilities are shown together with the sample proportions. Only the MNL quadratic form is chosen for this comparison, where it is sufficient to illustrate the differences in the level of precision. The Delta method is used to estimate the pointwise standard errors in the MNL, and the bootstrap method is used to estimate the standard error of the kernel estimated conditional probabilities, as recommended by Li and Racine (2003) and Li and Racine (2007). The confidence bands of the kernel estimates are wider in the ranges when few sample points are observed, and in those ranges the probability is either under-estimated or over-estimated. Then the confidence bands properly capture the information from the sample and indicate what level at which each pointwise predicted conditional probability should be trusted. This data density property is ignored by the MNL, which still suggests a high level of precision even at the ranges when there are no sample observations. In Figure 2.3 the MNL predicted conditional probability of females with low qualification and no children is the highest at ages 38-40 while the kernel estimates and the sample proportions show a decrease in the probability at this range. The worst performance of the MNL model is in predicting the propensity to work for the females with three or more children. In reverse to the MNL, the kernel estimate shows an outstanding performance in all sample cells.

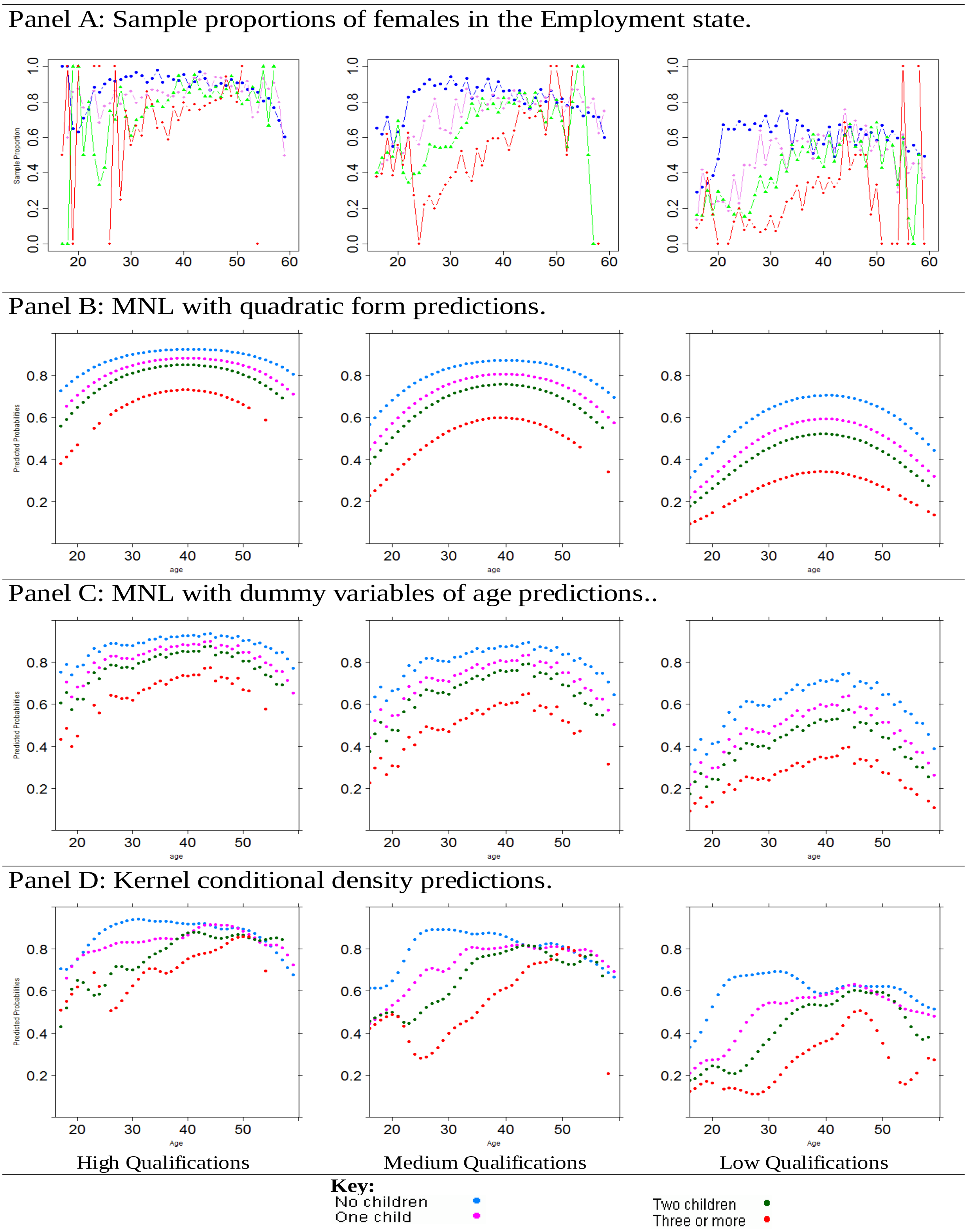


Figure 2.1: Predicted probability of being employed against age, restricted MNL model.

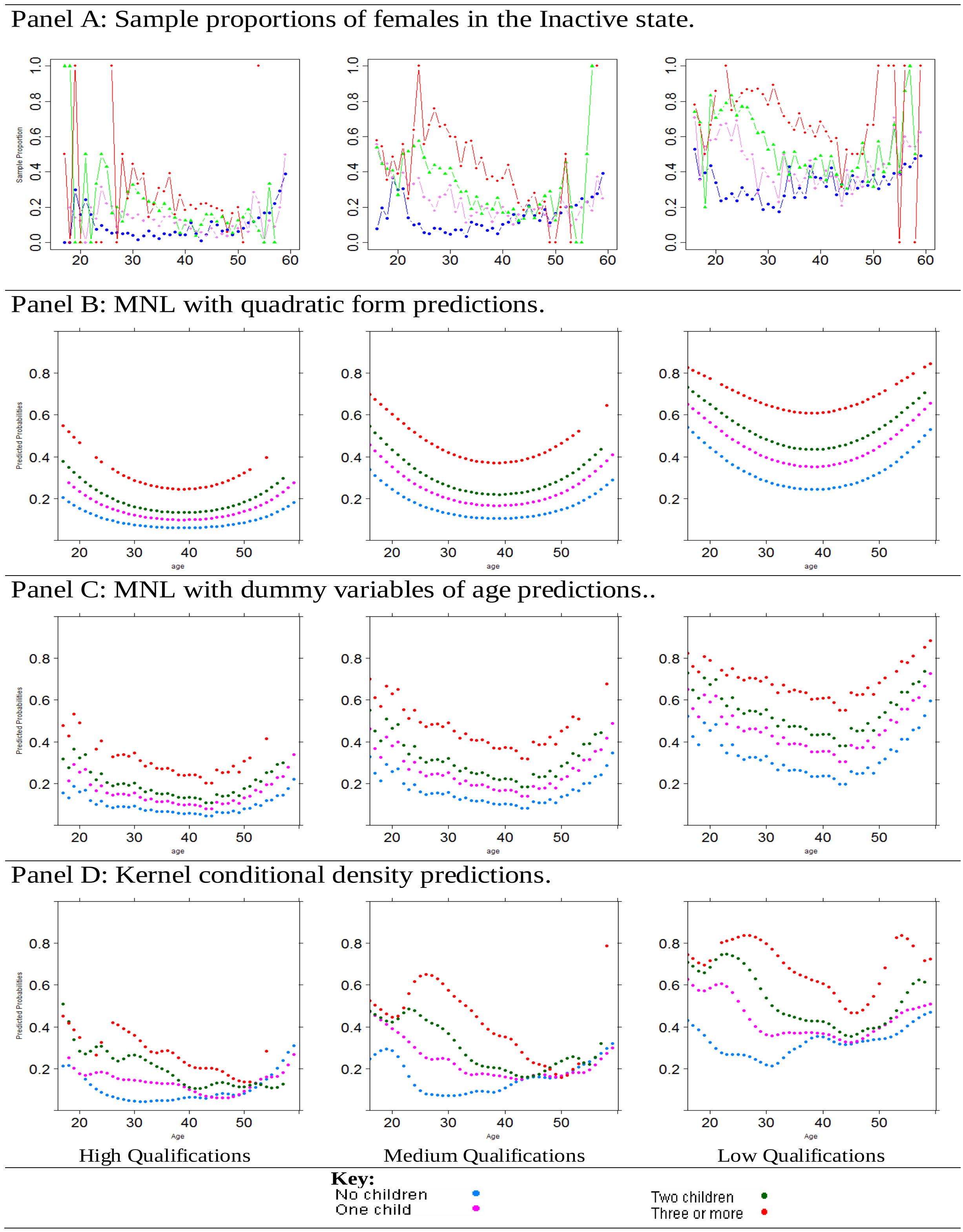


Figure 2.2: Predicted probability of being inactive against age, restricted MNL model.

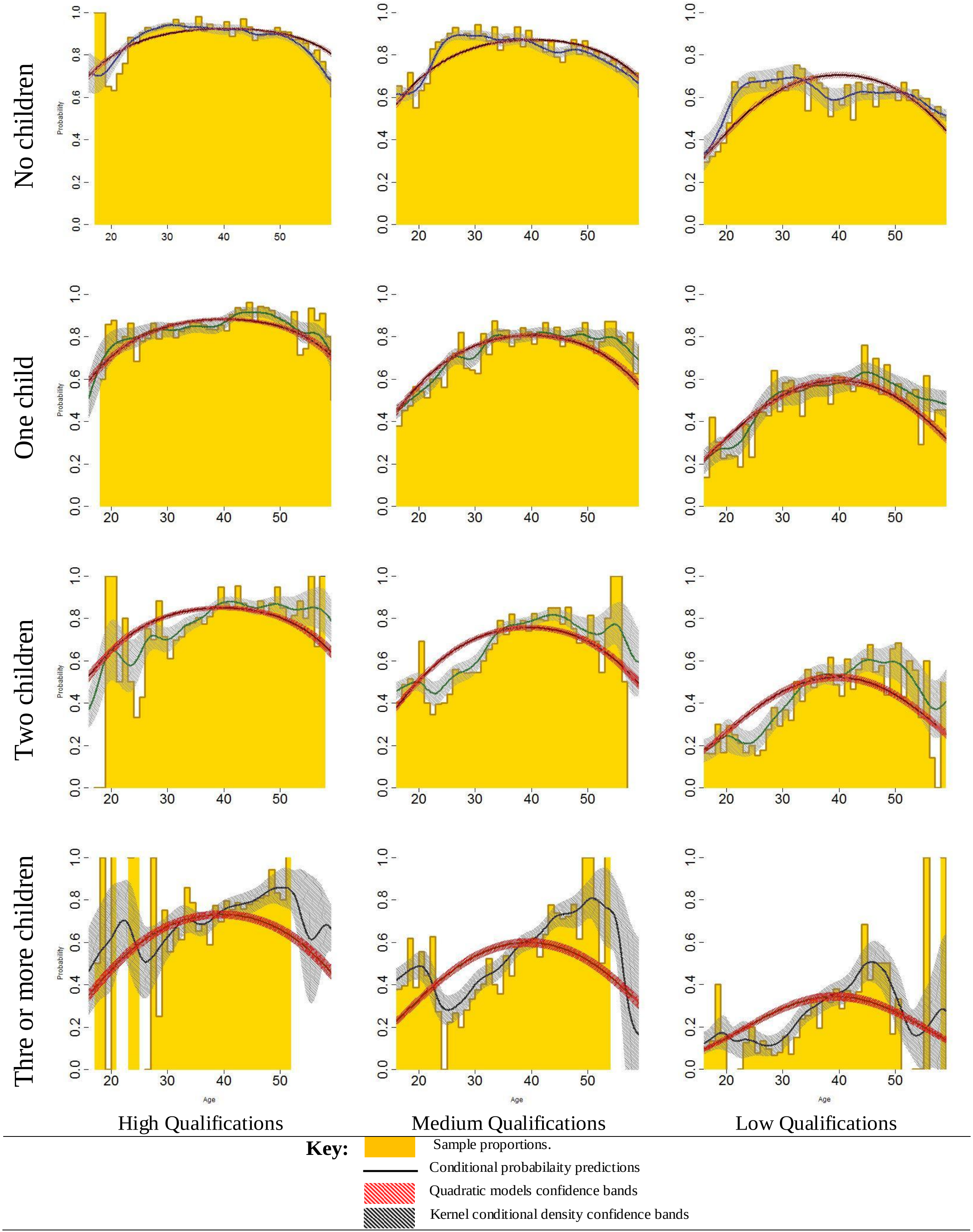


Figure 2.3: Predicted probabilities with the 95% confidence bands for being employed alternative from the MNL with quadratic form model and the kernel conditional density.

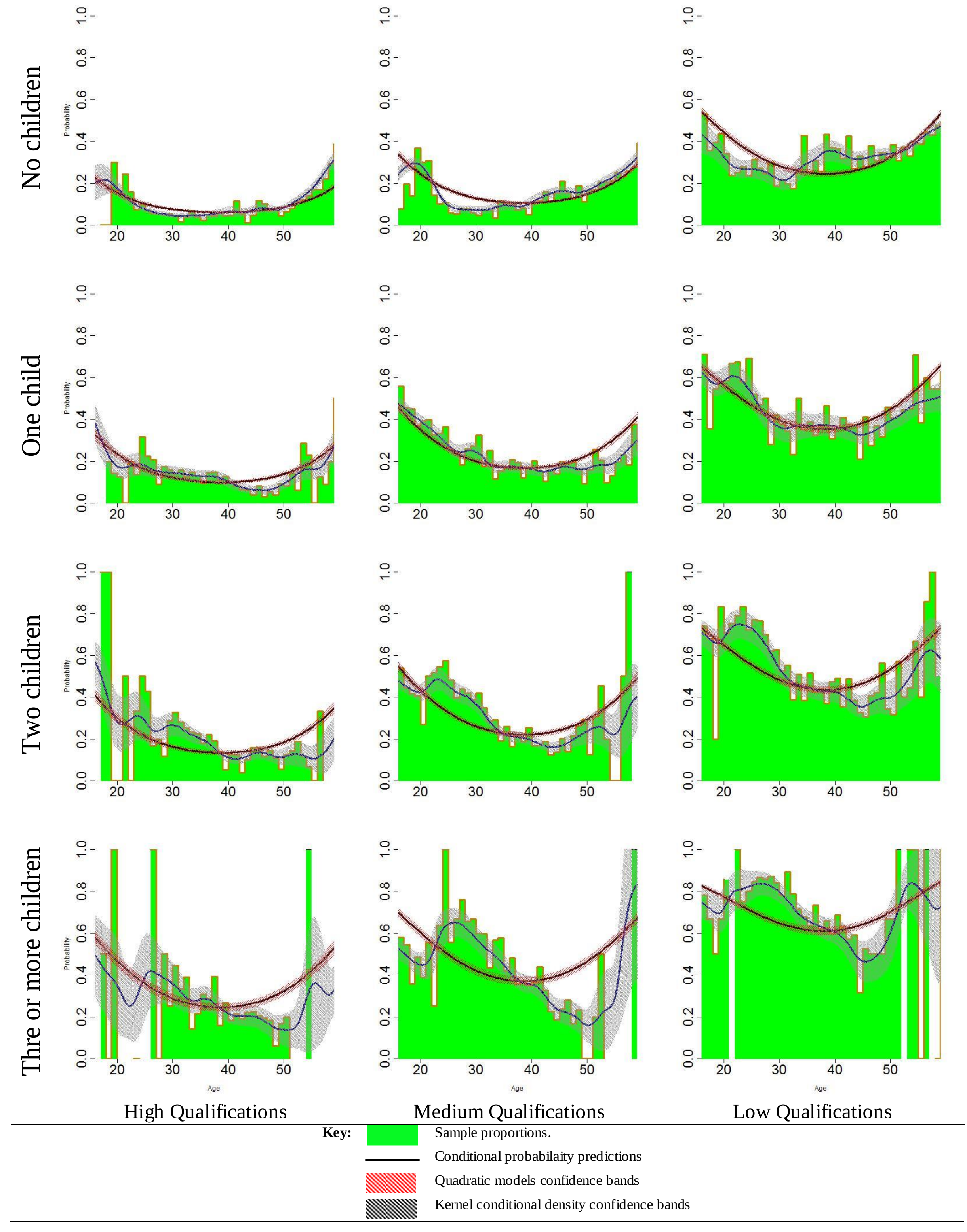


Figure 2.4: Predicted probabilities with the 95% confidence bands for being inactive alternative from the MNL with quadratic form model and the kernel conditional density.

2.5 The Unrestricted Models

The discussion of the restricted models demonstrates that the outcome of the MNL model deteriorates when defining a restrictive parametric functional forms to capture the effect of the continuous variables. Kernel conditional density estimates produce substantially better and reliable results, that match female labour force participation in the presented literature in Section 3, than the MNL model. The kernel estimates show that the age at which the propensity to work of females reaches the highest level is not the same for females with different number of dependent children and different eduction levels. The restricted model, however, use a smaller number of independent variables, so it explains less heterogeneity than the unrestricted model that are used in this section.

In this section we extend the specification of the functional form of the parametric model in an attempt to make the MNL model produce results similar to that shown by the kernel estimate. This may reduce the differences in the predicted probabilities between the MNL estimator and produce results that much better. Li and Racine (2007) claim that a correctly specified functional form, if available, is preferred to a nonparametric estimator, since a correctly specified parametric model is easier to estimate. In this research, in the following sections, we examine different specifications for the discrete choice model in an attempt to find the right functional form compared with the kernel nonparametric estimator.

The MNL model in the restricted model analysis shows improvement when the restrictive parametric quadratic form is relaxed to a semi-parametric form. So the new MNL specifications are developed in this direction, where more dummy variables are added in the semi-parametric form of the model. This on the other hand means that more coefficients are added in the model. The new dummy variables that are added are interaction terms of the age dummy variables with the dummy variables of the categorical variables. The number of the new interaction dummy variables is large since there are 42 year of age specific dummy variables. In the MNL, because the coefficients are alternative specific, the interaction of the age dummy variables with each of the other dummies generate 84 new coefficients. This makes the new models (coefficient rich) models.

The interaction terms (the slope coefficients) are added to the number of dependent children and education level dummy variables first. In the models that are estimated in the sub-samples, however, the interaction terms are added to the number of dependent children dummies only, since all the females in the sub-sample are in the same education level. The number of the new coefficients, from the quadratic form when the interaction terms are included, exceed 500 coefficients in the sample of all females and 300 coefficients in the sub-samples. The model with the interactions of age with number of children and education level is denoted by **age dummies with interactions 1** form. The second interactions model includes, in addition to the interaction of age with number of children and education level, the interaction of age with head of household, X^d_{hoh}. To avoid a possible perfect multicollinearity problem no interaction terms are added to age dummy variables with under 5 dependent children, X^d_{u5c}, since this is very likely to be correlated with the interactions of age with number of dependent children. This model is denoted by **age dummies with interactions 2** form. The LR test in Tables 2.5 and 2.4 show that the interaction term models are accepted against the quadratic form model.

Tables 2.5 and 2.4 show Likelihood Ratio test results and the degrees of freedom of the test for the unrestricted model. Appendix C shows the functional form of each model. The null hypothesis is the the coefficients that are added to the model are not significant, so rejecting the null hypothesis means that the form that is tested against the linear form is rejected. The panels in each table show the test of the linear form in the null against the quadratic form, the age dummies form and the interaction dummies forms. The MNL models are all strongly accepted against the linear form, so the linear form is not used in any further analysis.

Table 2.4: LR test of the MNL models[1]

		All females	**sub-samples**		
			High Qual.	GCE/GCSE	Low Qual
Linear age form	LR	1390.3481	244.2615	740.7486	277.0995
vs	df	2	2	2	2
Quadratic age form	p-value	0.0000	0.0000	0.0000	0.0000
Linear age form	LR	1544.6526	335.5853	877.9067	408.0022
vs	df	82	82	84	84
Age dummies form	p-value	0.0000	0.0000	0.0000	0.0000
Linear age form	LR	807.2954	262.7919	471.9074	276.6197
vs	df	424	226	244	250
Age dummies with interactions 1 form	p-value	0.0000	0.0469	0.0000	0.1190
Linear age form	LR	1008.0401	382.7069	635.7328	389.8687
vs	df	510	308	330	336
Age dummies with interactions 2 form	p-value	0.0000	0.0024	0.0000	0.0227

[1] Functional form of the MNL models is presented in Appendix C.

Table 2.5: LR test of the BNL models[1]

		All females	sub-samples		
			High Qual.	GCE/GCSE	Low Qual
Linear age form	LR	63.9777	12.7415	37.6773	19.5673
vs	df	1	1	1	1
Quadratic age form	p-value	0.0000	0.0004	0.0000	0.0000
Quadratic age form	LR	131.5605	53.3878	92.9801	90.0488
vs	df	42	40	42	42
Linear age form	p-value	0.0000	0.0930	0.0000	0.0000
Quadratic age form	LR	285.1997	132.1117	150.5406	185.5728
vs	df	209	108	119	159
Age dummies with interactions 1 form	p-value	0.0004	0.0574	0.0268	0.0734
Linear age form	LR	328.6070	197.2187	192.4168	158.4192
vs	df	252	149	162	161
Age dummies with interactions 2 form	p-value	0.0008	0.0050	0.0515	0.5428

[1] Functional form of the BNL models is presented in Appendix C.

The marginal effects for the quadratic form and the semi-parametric form models are reported in Table 2.6. The marginal effects of the BNL are shown in the last column of the table. The effect of the new variables, the head of household dummy and under 5 years old dummy, on the estimated probabilities is highly distinguishable. Females with under 5 years old child/children have 17-18 percentage points lower conditional probability of being employed than females with no under 5 dependent children. This exceeds the drop in the propensity to work that is generated by the 2 dependent children dummy variable. The effect of the head of household dummy on the probability of being in the employed state is negative, but the estimated drop in the estimated conditional probability is only 10 percentage points. The marginal effects of the quadratic form model and semi-parametric form on the sub-samples are presented in the Appendix C. The marginal effects of the interaction models are not reported for brevity.

The results of kernel conditional density estimator are reported in Table 2.7. The estimated MLCV bandwidths in the table indicate that the variables in the model are all relevant. However, the bandwidth of X^d_{u5c} is estimated very low in the multinomial conditional density in the sample of all females, and in the sub-sample of females with high qualifications. In the binomial conditional density kernel estimate on the other hand, the estimated MLCV bandwidth is exactly 0.5, except in the sub-samples of females with high qualification. This is possibly due to that there are very few females have children under 5 years old in the sub-samples where the bandwidth of X^d_{u5c} is estimated equals 0.5.

If a female has an under 5 years old child(ren) she is more likely to be out of the labour force, economically inactive, rather than being unemployed. So, $X^d_{u5c} = 1$, a female has under 5 child(ren), is not observed sufficiently for the unemployment alternative. On the other hand, X^d_{u5c} is correlated with X^d_{nc}, where $X^d_{u5c} = 1$ only if $X^d_{nc} > 0$. For the females in the sample with all dependent children younger than 5 the variable X^d_{u5c} adds little information. This is likely to happen for females in the age groups (16-50), whilst females at higher age group (50-59) are more likely not to have under 5 dependants. Female with high qualifications in the sample have less tendency to be inactive, the probability of this alternative is almost zero in the qualifications category. Females with dependent children are observed in the age group (20-50) and their probability of being inactive is lower than females with other qualifications. This is captured by the MNL model and kernel estimator jointly. For the kernel estimator the lowest plots in Figure 2.5 and 2.6 show a remarkable decrease in the predicted probability of inactive alternative that are estimated by the kernel estimator for the females with high qualifications. Then for females with high qualification $X^d_{u5c} = 1$ is almost not observed at the inactive alternative.

In contrast to this in the binary model $X^d_{u5c} = 1$ is almost not observed for the females with medium and low qualification. Females in this groups tend to be economically inactive if they have dependent children. This behaviour of X^d_{u5c} with both X^d_{nc} and Y^d affects the estimated bandwidths of the latter. As the estimation of the bandwidths of kernel fixed bandwidths estimators is affected by the extreme values outliers in the data in the continuous variables. In the case of smoothing a number of discrete variables, like in the above, a problem may exist in the estimated bandwidths, which we should argue that it should have closed look in research.

Figures 2.5 to 2.12 show the predicted conditional probabilities of the: MNL model with quadratic form in age, the MNL model with age dummy variables, the MNL model with age dummy variables with interactions 1, the MNL model with age dummy variables with interactions 2 and the kernel conditional density estimates, respectively from top to bottom. All the figures show that it is very difficult to draw a conclusion from the plots of the models with the interaction terms. It seems that adding coefficients increases the noise in the model. For some interaction dummy variables there are very few number of observations in the sample. This makes the estimated coefficients imprecise. Increasing the number of variables in a parametric model increases the possibility of multicollinearity problem and the singularity in the data matrix. This is likely to be the problem with the models with the interaction terms.

Kernel conditional density predicted probabilities show outstanding performance, similarly to that shown in the restricted model. The propensity to work for the females is estimated consistently with the suggestions in the literature and female labour force participation theory. Females with more dependent children and low qualification level have less conditional probability to be in the employed state or be in the labour force. The number of children under 5 years old has a substantial effect in reducing the propensity of females to work. After the work premium age, after 40 years old, the propensity to work of the females with dependent children become almost similar to the propensity of the females with no dependent children. Note that the propensity to work for females with no dependent children is high at all age groups. Then the differences in the propensity to work decrease due to the increase of the propensity of the females with dependent children. However, Figures 2.6 and 2.8 show that females with children under 5 years old and in the work premium age are still having a substantially lower propensity than other females.

Table 2.6: Marginal effects for the unrestricted MNL and the unrestricted BNL models[1].

		MNL			BNL[4]
		Employed	Unemployed	Inactive	Unemployed
Panel A: Models with quadratic form in X^c_{age}					
$p\left((\cdot)^d \,\middle\vert\, \overline{\mathbf{X}}\right)^2$		0.733	0.033	0.234	0.037
Age($= X^c_{age}$)		5.128 (0.140)	-0.387 (0.049)	-4.741 (0.132)	-0.650 (0.057)
Age Squared		-6.588 (0.188)	0.322 (0.068)	6.266 (0.177)	0.640 (0.078)
Number of Children[3] X^d_{nc}:					
	1 child	-0.072 (0.008)	0.006 (0.003)	0.066 (0.007)	0.013 (0.003)
	2 children	-0.140 (0.009)	0.001 (0.003)	0.139 (0.008)	0.011 (0.004)
	3 or more	-0.304 (0.012)	0.008 (0.004)	0.296 (0.012)	0.037 (0.007)
X^d_{hoh}		-0.103 (0.006)	0.020 (0.003)	0.083 (0.006)	0.031 (0.003)
X^d_{u5c}		-0.173 (0.008)	-0.007 (0.002)	0.180 (0.008)	-0.0001 (0.003)
Education Level[3] X^d_{ed}:					
	ges/gcse	0.193 (0.006)	-0.023 (0.002)	-0.170 (0.005)	-0.034 (0.002)
	high qual.	0.267 (0.005)	-0.028 (0.002)	-0.239 (0.005)	-0.044 (0.002)
Log likelihood		-22474.156			-4737.79
Pseudo R^2		0.1223			0.0965
Wald stat.		5218.27			908.14
Panel B: MNL with year dummies of X^c_{age}					
$p\left((\cdot)^d \,\middle\vert\, \overline{\mathbf{X}}\right)^2$		0.734	0.033	0.233	0.037
Age dummies			the marginal effects of the age dummies are not shown		
Number of Children[3] X^d_{nc}:					
	1 child	-0.060 (0.008)	0.002 (0.003)	0.058 (0.008)	0.007 (0.003)
	2 children	-0.124 (0.009)	-0.003 (0.003)	0.127 (0.009)	0.004 (0.004)
	3 or more	-0.283 (0.013)	0.003 (0.004)	0.280 (0.013)	0.025 (0.007)
X^d_{hoh}		-0.105 (0.006)	0.022 (0.003)	0.085 (0.006)	0.033 (0.003)
X^d_{u5c}		-0.189 (0.010)	-0.003 (0.003)	0.192 (0.010)	0.006 (0.004)
Education Level[3] X^d_{ed}:					
	ges/gcse	0.195 (0.006)	-0.023 (0.002)	-0.172 (0.005)	-0.034 (0.003)
	high qual.	0.265 (0.005)	-0.028 (0.002)	-0.239 (0.005)	-0.042 (0.002)
Log likelihood		-22397.00			-4703.9989
Pseudo R^2		0.1253			0.103
Wald stat.		5313.32			1035.21
n		35311			26007

[1] Unrestricted models add X^d_{hoh} and X^d_{u5c} dummy variables to the restricted models in Table 2.2 .
[2] $p\left((\cdot)^d \,\middle\vert\, \overline{\mathbf{X}}\right)$ indicates $p\left(y^d \,\middle\vert\, \overline{\mathbf{X}}\right)$ in the MNL model and $p\left(u^d \,\middle\vert\, \overline{\mathbf{X}}\right)$ in the BNL model.
[3] The basegroups are: for the Number of Children, the no children group; for the Education level, the Low Qualification group; for X^d_{u5c}, no under 5 children; X^d_{hoh}, not head of the household.
[4] The marginal effects of the BNL are for the probability of unemployment for the females in the labour force.

Table 2.7: Bandwidths of the unrestricted kernel nonparametric conditional pdf models for for all females and the labour force subsample[1].

		Highest Qualification sub-samples		
	All females	High Qual.	GCE/GCSE	Low Qual
Panel A: Multinomial Kernel nonparametric conditional density[2]				
Eco. activity Y^d	0.004172	0.006167	0.004722	0.002627
Age X^c_{age}	1.960021	1.691900	1.716179	2.532146
Number of Children X^d_{nc}	0.163338	0.202886	0.129685	0.195695
Head of household X^d_{hoh}	0.069432	0.2316024	0.038003	0.086186
Under 5 children X^d_{u5c}	6.92e-16	2.58e-14	0.035466	0.015003
Eduction level X^d_{ed}	0.030360			
CV. Optimum Value	0.631176	0.446865	0.653140	0.798135
log likelihood	-21900.44	-4420.153	-10402.42	-7079.773
CV iterations	1	5	2	2
n	35311	10125	16181	9005
Panel B: Binomial Kernel nonparametric conditional density[2]				
Eco. activity Y^d	0.001533	0.001703	0.001628	0.001668
Age X^c_{age}	2.030528	2.252071	2.055114	1.972357
Number of Children X^d_{nc}	0.365660	0.189327	0.278545	0.712765
Head of household X^d_{hoh}	0.030499	0.269757	0.008774	0.040048
Under 5 children X^d_{u5c}	0.5	0.243203	0.5	0.5
Eduction level X^d_{ed}	0.027529			
CV. Optimum Value	0.181522	0.113285	0.191092	0.278112
log likelihood	-4614.677	-974.8996	-2271.539	-1368.491
CV iterations	1	5	5	5
n	26007	8869	12111	5027

[1] The bandwidth are estimated using the MLCV method.

[2] The kernel functions are: Second Order Gaussian kernel for the continuous variables, Aitchison and Aitken (1979) kernel function for the unordered discrete variables and Geometrical kernel function for the ordered discrete variables.

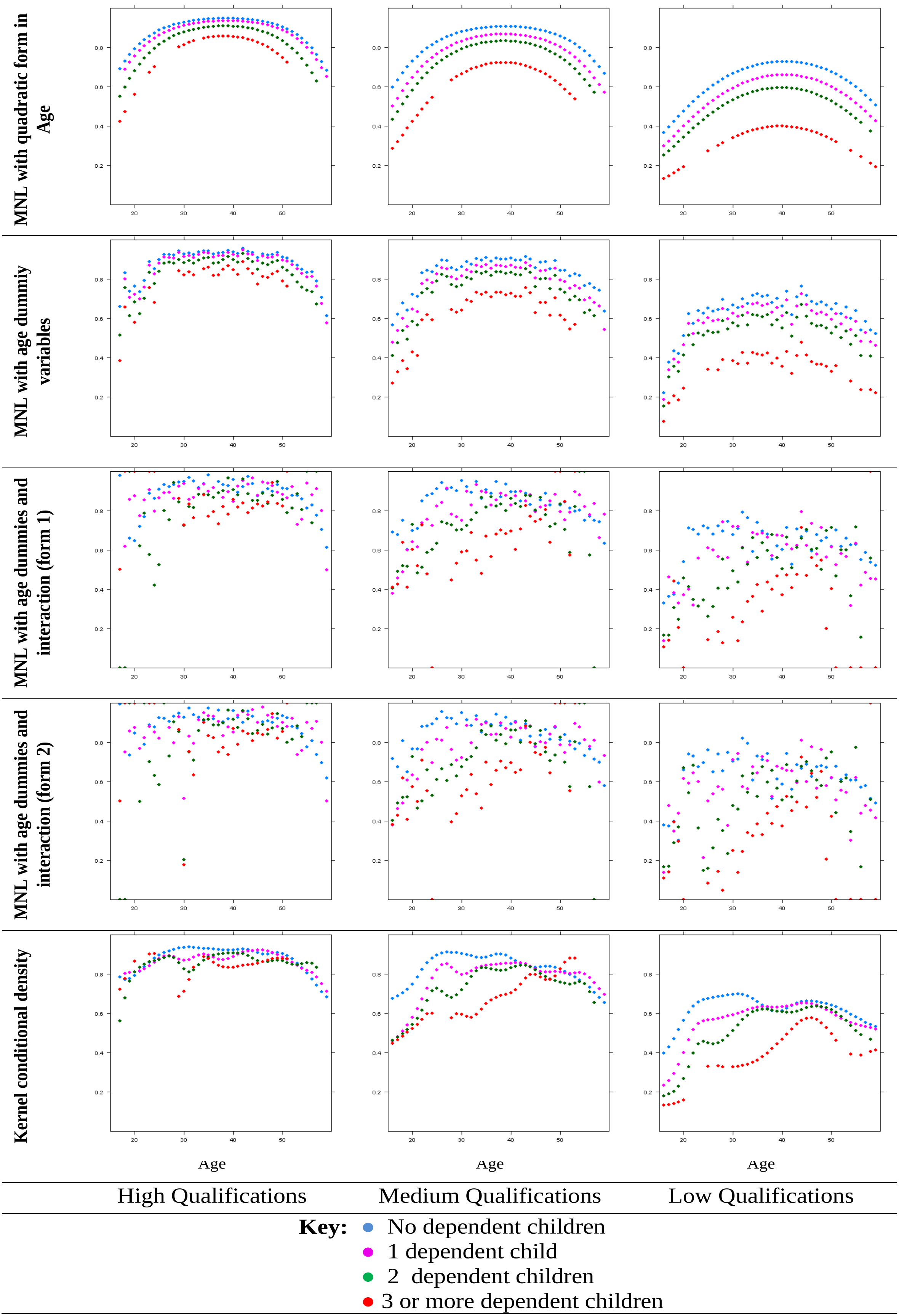


Figure 2.5: Probability of employment for the females who are not head of household and not have under 5 years old children.

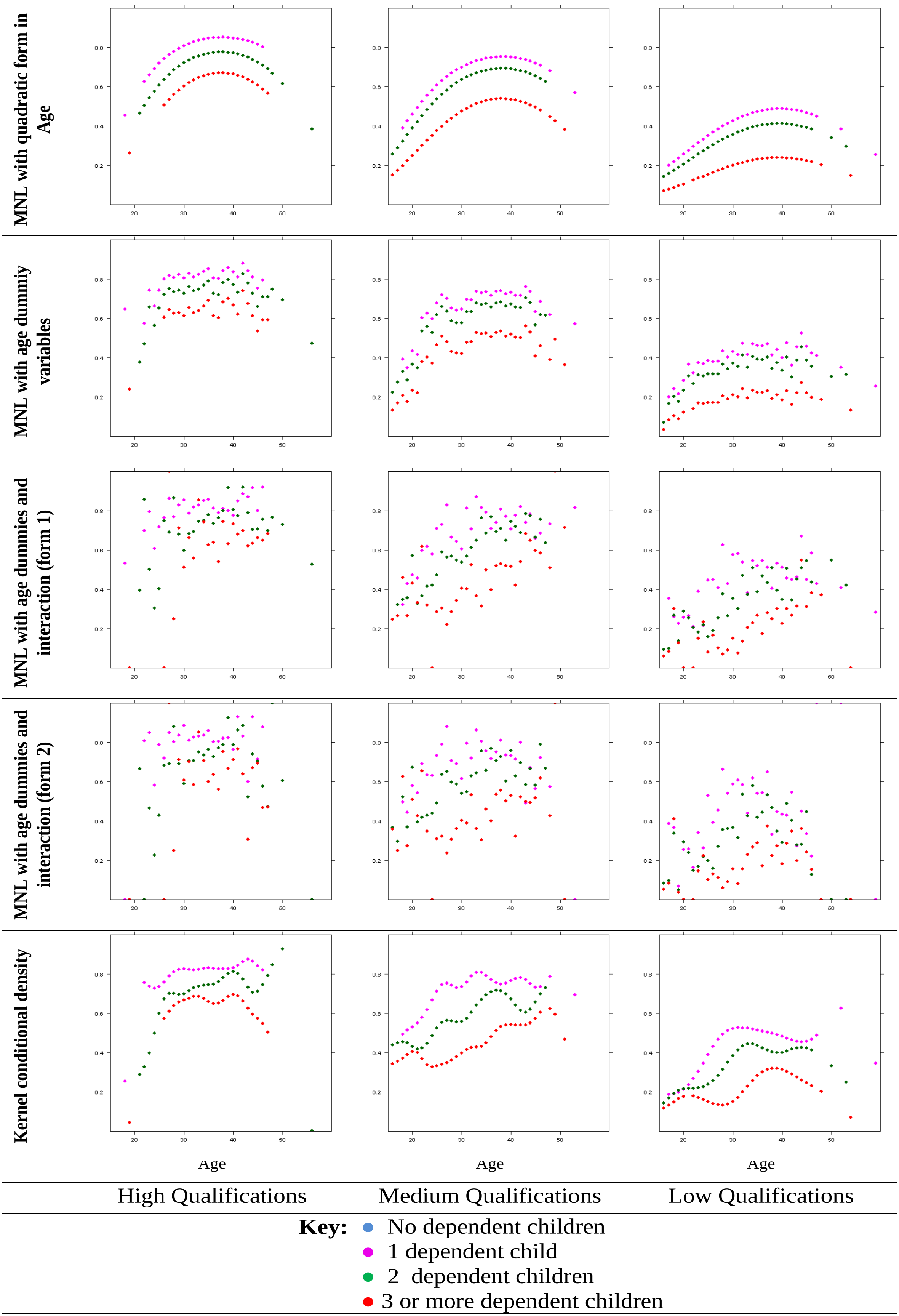


Figure 2.6: Probability of employment for the females who are not head of household and have under 5 years old children.

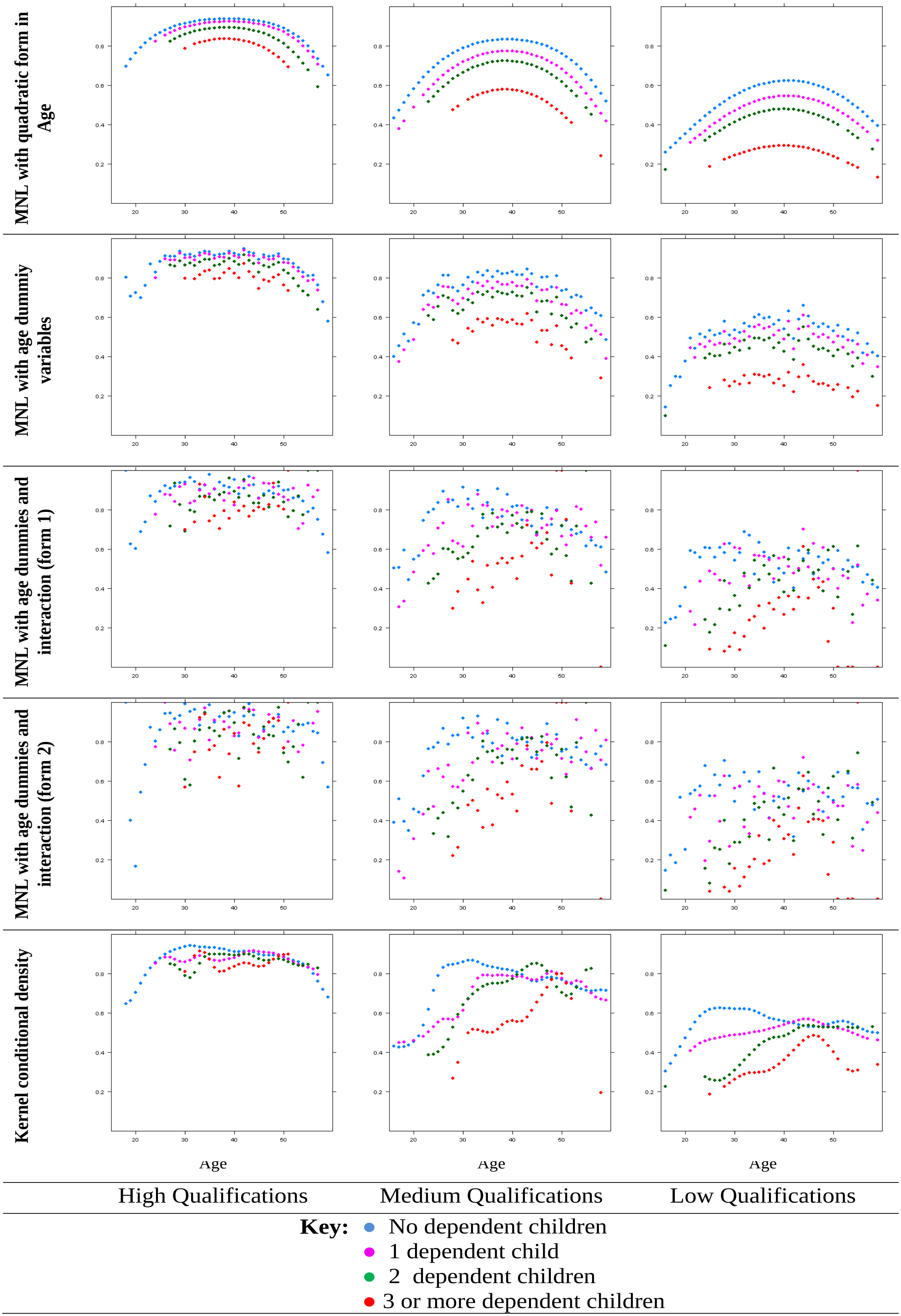


Figure 2.7: Probability of employment for the females who are head of household and have no under 5 years old children.

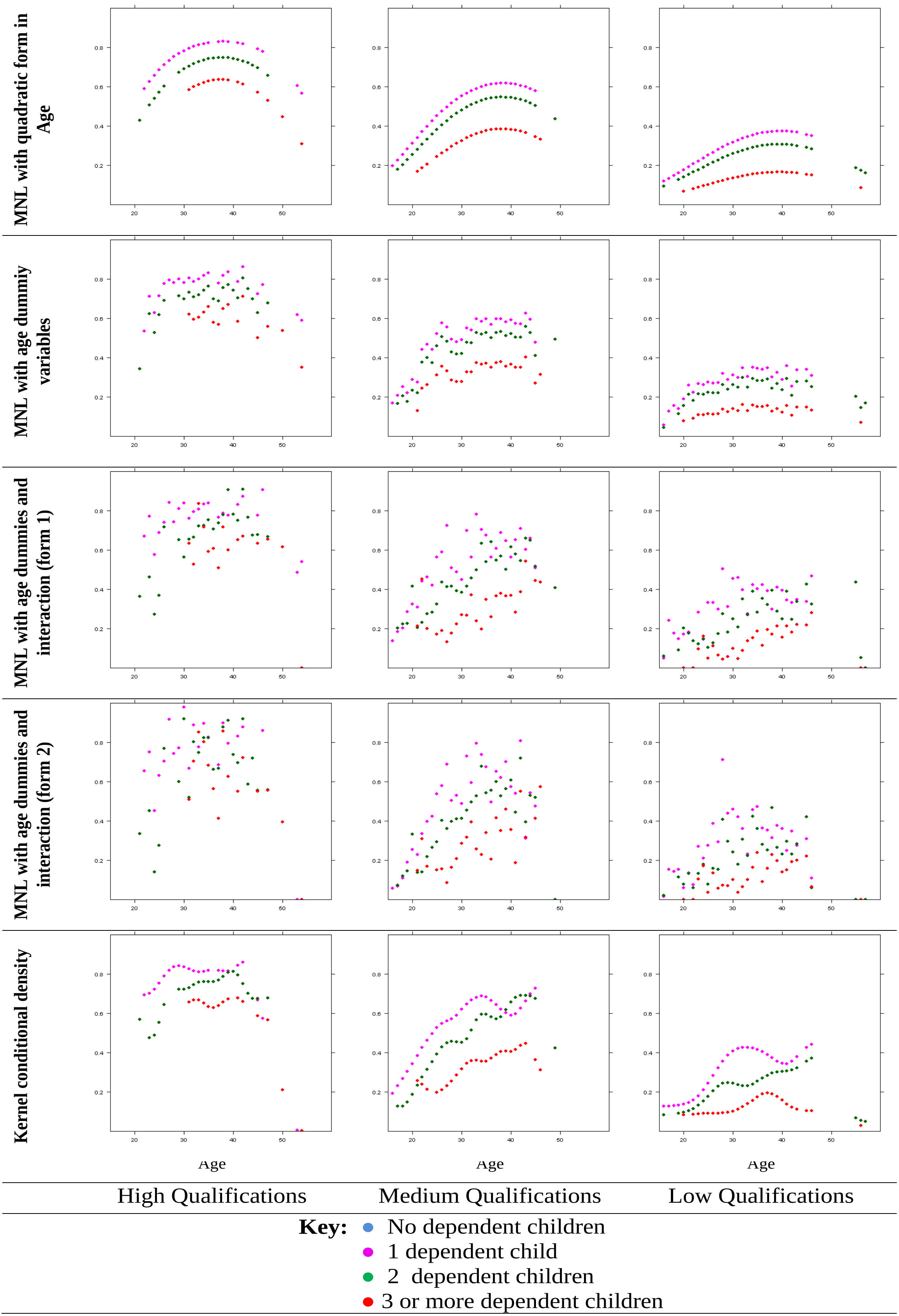


Figure 2.8: Probability of employment for the females who are head of household and have under 5 years old children.

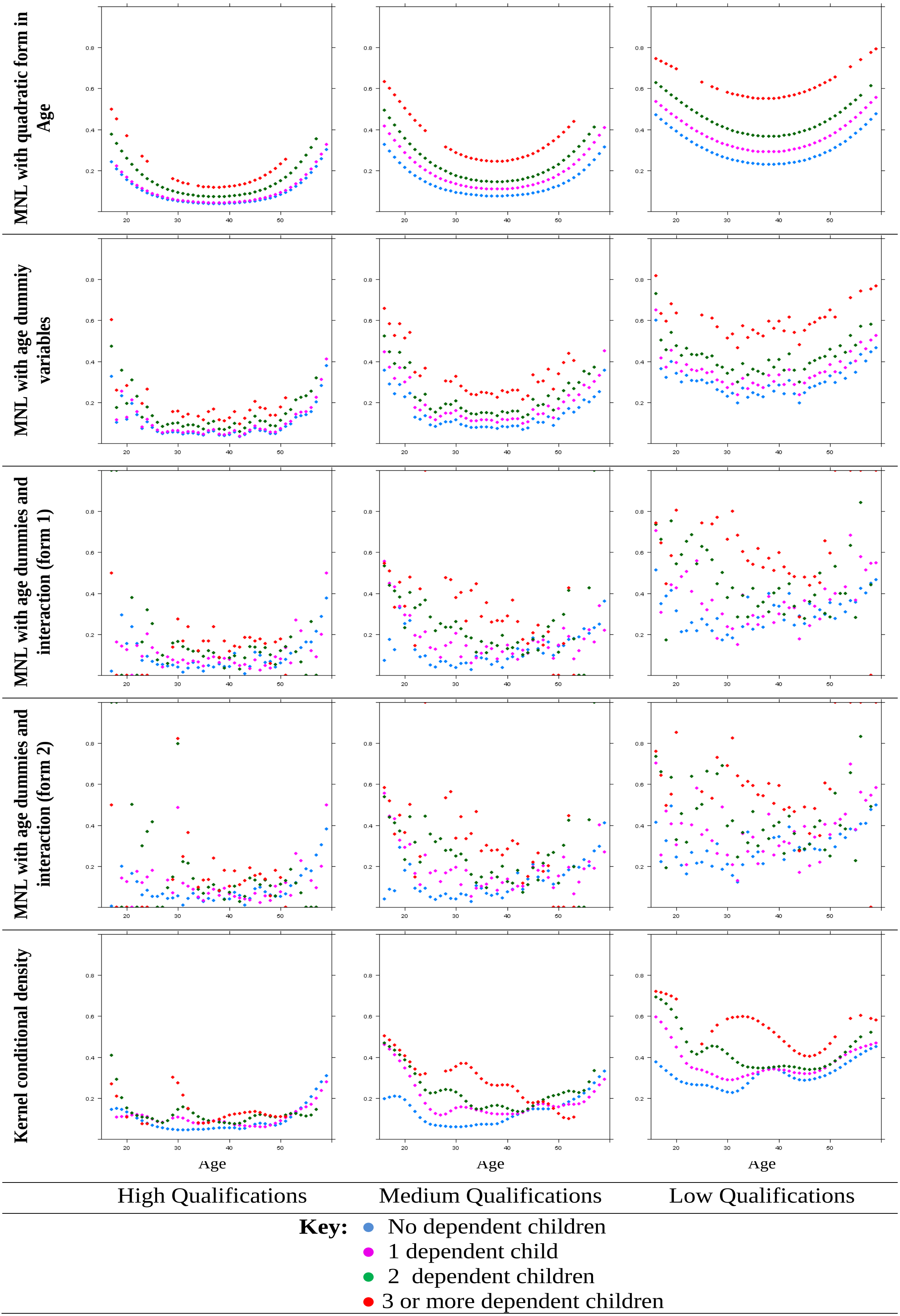


Figure 2.9: Probability of inactive for the females who are not head of household and have no under 5 years old children.

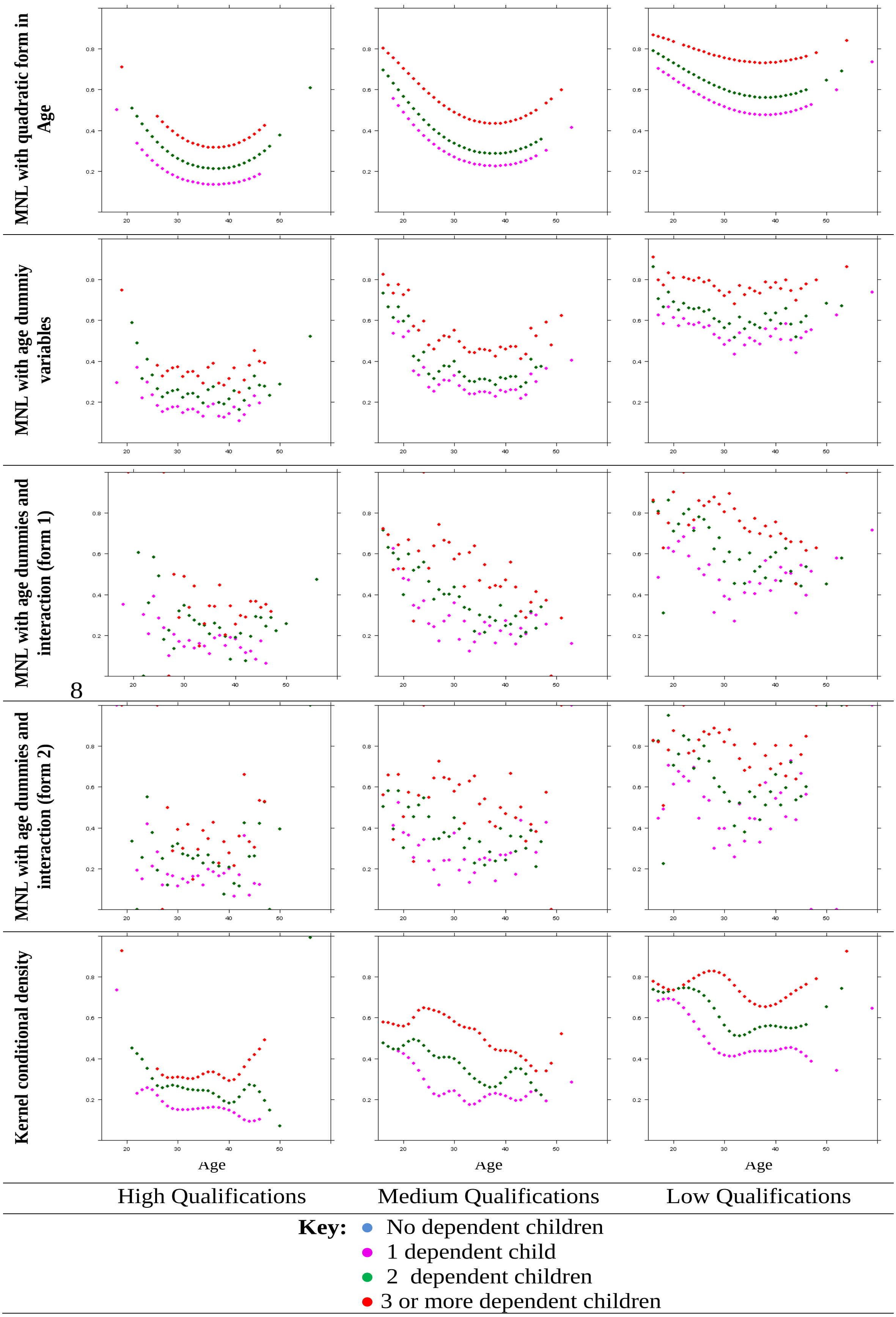


Figure 2.10: Probability of inactive for the females who are not head of household and have under 5 years old children.

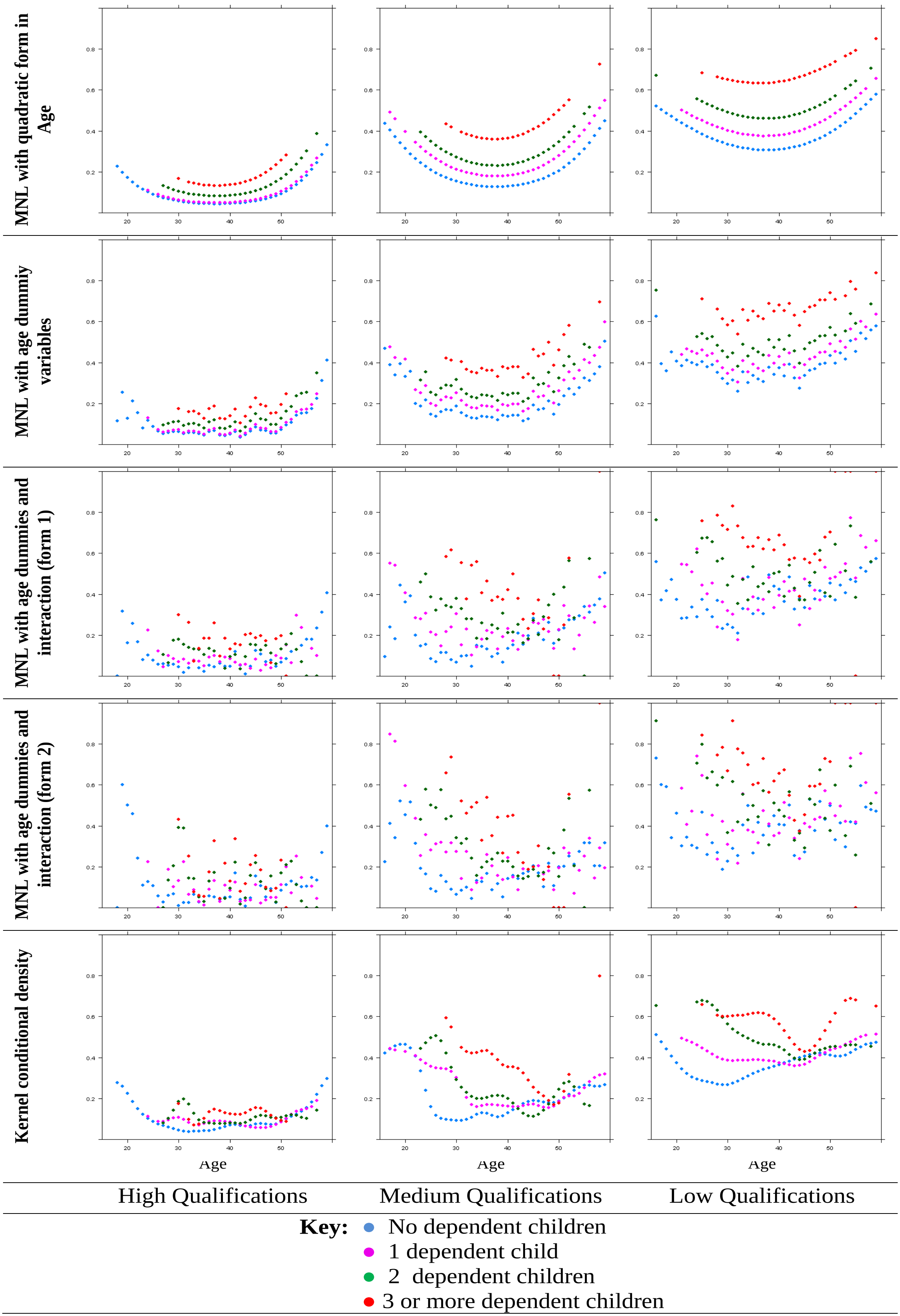


Figure 2.11: Probability of inactive for the females who are head of household and have no under 5 years old children.

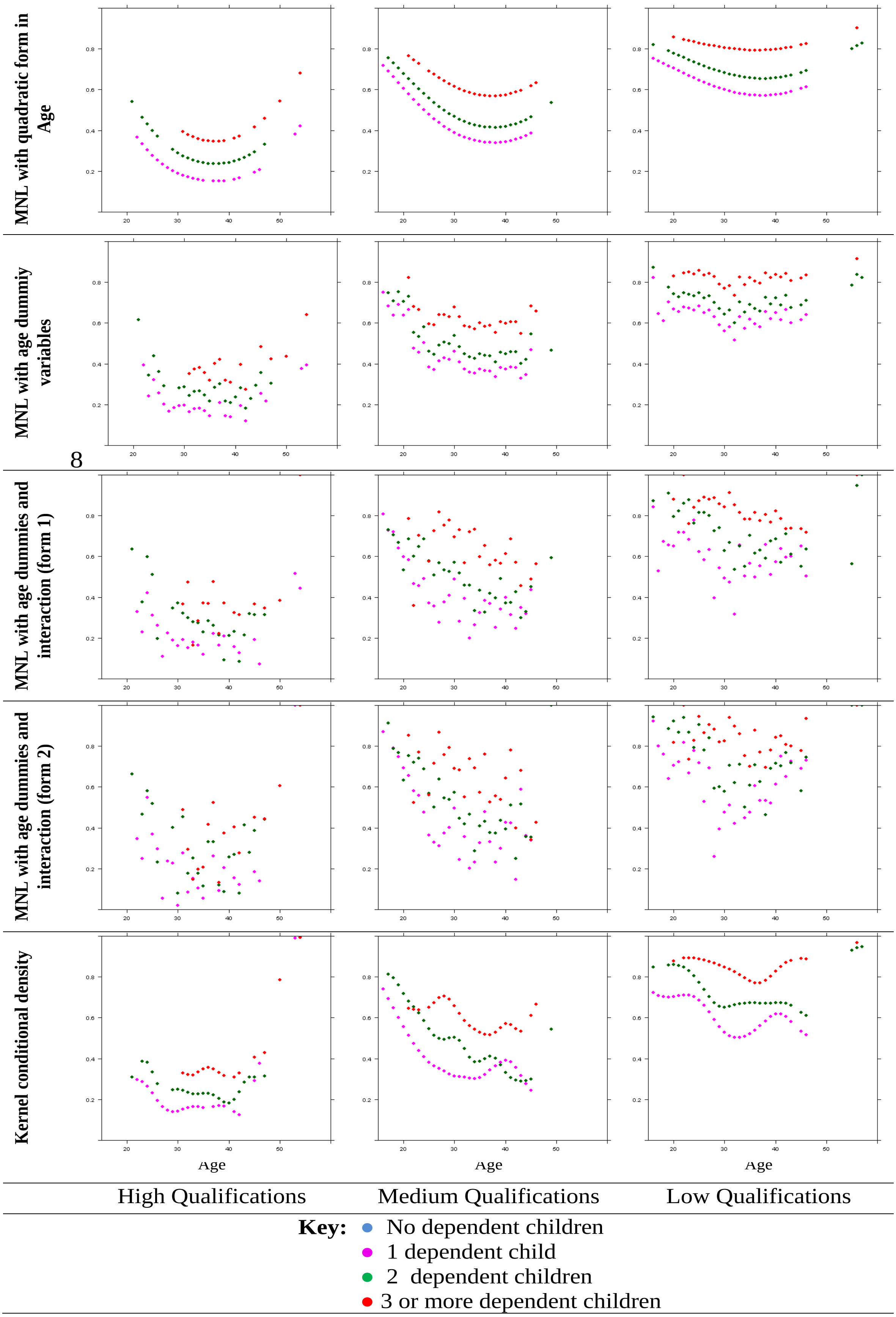


Figure 2.12: Probability of inactive for the females who are head of household and have under 5 years old children.

2.5.1 The Nonparametric Tests

Kernel Smoothing Specification tests

The correct specification test for the parametric models are performed using the MLCV bandwidths of the kernel conditional density estimate. The results of the tests are shown in Tables 2.8 and 2.9. Under the null hypothesis that the test statistic converges to a standard normal. If the null hypothesis is not true, the test statistics converges in probability to a positive number. The distribution of the test statistic is approximated empirically using the bootstrap method that was described in Section 2.3.3. The number of bootstrap samples is 399. The test is one side tail test. The p-value that are reported in the table shows that the empirical distribution of the test statistics drifts from that of a normal distribution. The sign and the empirical distribution of the test statistic depends crucially on the term $\frac{I\left(y_i^*=y_j^*\right)-p\left(y_i^*|x_j,\widehat{\boldsymbol{\beta}}^*\right)}{p\left(y_i^*|x_i,\widehat{\boldsymbol{\beta}}^*\right)}$. The test statistic can be negative in the case of a rare events, like in the case of the number of females in the unemployment alternative, $p(U^d=1)$ and $p(Y^d=1)$. This case is not considered in any of the papers that developed these tests. Then the bootstrap empirical distribution drifts to the left, but this may not affect the test decision where the null is not true and the distribution is different than standard normal.

The Zhang test and Fan test produce similar results regarding the correct specification form of the MNL models. The quadratic form MNL and the semi-parametric form in age are both highly rejected in all samples. The MNL model with the interaction terms 1 model on the other hand, is accepted in Zhang test in the sample of all females and the sub-sample of females with medium qualifications. The MNL with interactions 2 fails in the sample of all females. Both tests require very hard calculations, therefore, it becomes very difficult to approximate the bootstrap empirical distribution due to the large sample size. The p-value of the tests in the sub-samples indicate less difference between the parametric MNL model with interactions 2 and the kernel conditional density model. The MNL with interactions 2 specification failed to reject in all sub-samples, except the sub-sample of females with low qualifications.

The result of the BNL specification tests are very weak compared with the results of the tests of the MNL models. A problem appeared with BNL specification tests, that is generated from the distance term $\frac{I\left(y_i^*=y_j^*\right)-p\left(y_i^*|x_j,\widehat{\boldsymbol{\beta}}^*\right)}{p\left(y_i^*|x_i,\widehat{\boldsymbol{\beta}}^*\right)}$. In most bootstrap samples this term is either zero or an undetermined $(\frac{0}{0})$ term. This is generated by that; $I\left(y_i^*=y_j^*\right)=p\left(y_i^*\,|x_j\,,\widehat{\boldsymbol{\beta}}^*\right)$ or $p\left(y_i^*\,|x_i\,,\widehat{\boldsymbol{\beta}}^*\right)=0$ in most bootstrap samples. The second case occurs more in the Fan, Li and Min (2006) test because the variance of the distance measure depends on the BNL the predicted probabilities. The tests seems to be working better in the case where there are no rare events in the model. The only results that are produced successfully are the results of the test of the BNL model with quadratic form in age. For other forms, the test statistic is calculated successfully for the semi-parametric form model but the bootstrap empirical distribution fails, while both test statistic and its empirical distribution are completely failed in the interaction form models.

The correct specification tests illustrate that; the quadratic, semi-parametric and the interactions 1 MNL forms are all rejected against the kernel multinomial conditional density estimates. However, as appeared in the plots above, the MNL with interactions 2 is parametrically an adverse specification which exhibits serious problems like multicollinearity. The conditional probabilities that are estimated by the interaction 2 models merely sensible. For the binomial conditional density our results allow to make decision about the quadratic model only. The quadratic form BNL model is rejected against the kernel estimate.

The discrete choice correct specification tests are slow, apply tough calculations and are difficult to manage. Despite that the logit model is an easy type of discrete choice models compared with multinomial probit, so that the bootstrap method and the simulation of the dependent variable to estimate the empirical distribution of the tests is easy theoretically, it has been difficult to approximate the distribution of the test statistics in many samples.

Table 2.8: Smoothing specification tests for the MNL models[1]

MNL Models[2]	All females[3]				High Qual sub-sample			
	Zhang Test		Fan Test		Zhang Test		Fan Test	
	Test Stat.	*Sig*	Test Stat.	*Sig*	Test Stat.	*Sig*	Test Stat.	*Sig*
MNL Model 1	97.831	*0.000*	42.963	*0.000*	4.708	*0.000*	2.732	*0.003*
MNL Model 2	84.373	*0.000*	37.007	*0.000*	1.9486	*0.000*	1.020	*0.000*
MNL Model 3	16.905	*0.644*	7.168	*0.000*	-8.9820	*0.000*	-4.950	*0.000*
MNL Model 4	na[a]	na[a]	na[a]	na[a]	20.495	*0.053*	-11.946	*0.138*

	Mid. Qual. sub-sample				Low Qual sub-sample			
	Zhang Test		Fan Test		Zhang Test		Fan Test	
	Test Stat.	*Sig*	Test Stat.	*Sig*	Test Stat.	*Sig*	Test Stat.	*Sig*
MNL Model 1	50.677	*0.000*	27.230	*0.000*	28.855	*0.000*	8.793	*0.000*
MNL Model 2	38.927	*0.000*	20.926	*0.000*	10.904	*0.000*	3.213	*0.000*
MNL Model 3	5.640	*0.065*	3.013	*0.017*	-8.364	*0.000*	-2.522	*0.000*
MNL Model 4	-15.411	*0.192*	-8.030	*0.247*	-23.632	*0.000*	-7.043	*0.006*

[1] Number of bootstrap samples is 399.
[2] MNL models specification is presented in Appendix C.
[3] na[a] denotes that the test is not performed for the sample of all females.

Table 2.9: Smoothing specification tests for the BNL models[1]

BNL Models[2]	All females[3]				High Qual sub-sample			
	Zhang Test		Fan Test		Zhang Test		Fan Test	
	Test Stat.	Sig	Test Stat.	Sig	Test Stat.	Sig	Test Stat.	Sig
BNL Model 1	12.456	0.000	8.078	0.000	0.890	0.050	0.586	0.048
BNL Model 2[4]	5.717	0.000	3.768	0.000	0.646	na^b	0.341	na^b
BNL Model 3[4]	-4.167	na^a	-2.520	na^a	na^b	na^b	na^b	na^b
BNL Model 4[4]	na^a	na^a	na^a	na^a	na^b	na^b	na^b	na^b
	Mid. Qual. sub-sample				Low Qual sub-sample			
	Zhang Test		Fan Test		Zhang Test		Fan Test	
	Test Stat.	Sig	Test Stat.	Sig	Test Stat.	Sig	Test Stat.	Sig
BNL Model 1	3.851	0.005	2.612	0.005	8.377	0.000	5.614	0.000
BNL Model 2[4]	-1.211	na^b	-0.819	na^b	-1.046	na^b	-0.698	na^b
BNL Model 3[4]	na^b	na^b	na^b	na^b	na^b	na^b	na^b	na^b
BNL Model 4[4]	na^b	na^b	na^b	na^b	na^b	na^b	na^b	na^b

[1] Number of bootstrap samples is 399.
[2] MNL models specification is presented in Appendix C.
[3] na^a denotes that the test is not performed for the sample of all females.
[4] na^b denotes that the empirical bootstrap distribution failed to estimate.

Tests for density equality

The kernel conditional density estimate predicts wide differences in the probability of being employed for females with different number of children and different qualifications levels. Our objective in this final investigation in this subsection is to examine the differences the conditional density of Y_i^d and U_i^d. In the kernel estimation framework there are methods to test the difference between conditional densities as well as between unconditional densities, using the densities estimated in the sub-sample and the cells of the categorical variables in the pooled sample. To apply the conditional density equality test we do need to write our own code in R, as we did to the discrete choice specification tests and the kernel estimators in the following chapters. Also, we do not need to replicate the simulations in Li et al. (2009), similarly to what we did to apply the discrete choice specification tests. This process is very long and difficult to manage within our PhD research period. However, the differences in the conditional density estimates imply differences in the joint and marginal densities of the variables in the model. So, our argument in this subsection is that testing the univariate unconditional densities in addition to the joint and marginal densities might provide information about the sources for differences in the estimated conditional densities.

Similarly to the correct specification smoothing tests, the bootstrap method is used to estimate the empirical distribution of the test statistic. The number of bootstrap samples is 399. The results are reported in Table 2.10. The unconditional kernel densities that are used in the test are; the univariate densities of the dependent variable, age, number of dependent children, under 5 years old children and head of household, in addition to the multivariate mixed type joint density and marginal density. The test is performed smoothly in all samples except for the joint density in the sample of all females.

Table 2.10 shows two tests for each unconditional density function; first a test that applied for the sample of all females, which uses the bandwidths that are estimated once in that sample, and test the equality of densities between different qualification levels. Which means that the categories of X_{ed} are used in the null hypothesis. If the unconditional density is the same in all the categories of X_{ed}, then the null hypothesis will be accepted in all three partial tests (high qualifications category v. medium qualifications category, high qualifications category v. medium qualifications category and medium qualifications category v. low qualifications category). Second, to test whether the unconditional kernel densities that are estimated in the sub-samples are the same. For the same variable three unconditional densities are estimated, one kernel unconditional density in each sub-sample, and the MLCV bandwidths are obtained to the test. To clarify this in the table we refer to the first test by the **same bandwidths test** and the second test as the **different bandwidths test**. The number of females is denoted by n_1, n_2 and n_3 for the females in high qualifications, medium qualifications and low qualification in each category/sub-sample.

The null hypothesis that the unconditional densities is the same is rejected in all samples, except for the unconditional density of X^d_{u5c} for the females with high qualification and the females with medium qualification. The standardised test statistics for the estimated density of the dependent variable, the estimated joint density and the estimated marginal density functions are in general higher for the tests between the group of females with high qualifications and the group of females with low qualifications, i.e. in the tests of $\widehat{f_1}(\cdot) = \widehat{f_3}(\cdot)$. The differences between the univariate unconditional densities estimates of the variables that included in the model cause the differences in the joint and the marginal densities estimates, since the later are multivariate unconditional densities that include those variables.

The differences that are captured by the unconditional densities equality tests are very likely to be affecting the conditional density estimates. Unconditional density equality tests show that Y^d, U^d, X^c_{age}, X^d_{nc}, X^d_{hoh} and X^d_{u5c}[2] are all having density functions that are dissimilar for females with different highest qualification levels. Those differences are influence on the conditional probability function kernel estimates and generate the different in the pointwise conditional probabilities that shown in the plots above.

[2]Except between females with high qualification and females with medium qualifications

Table 2.10: Unconditional densities equality tests

	Multinomial models sub-samples					
	$\widehat{f}_1(\cdot) = \widehat{f}_2(\cdot)$		$\widehat{f}_1(\cdot) = \widehat{f}_3(\cdot)$		$\widehat{f}_2(\cdot) = \widehat{f}_3()$	
	Test Stat	Sig.	Test Stat	Sig.	Test Stat	Sig.
Y^d						
Same bandwidths	202.37	0.000	1030.63	0.000	461.77	0.000
Different bandwidths	202.26	0.000	1030.50	0.000	461.93	0.000
X^c_{age}						
Same bandwidths	1365.63	0.000	141.31	0.000	-462.86	0.000
Different bandwidths	215.35	0.000	127.54	0.000	189.54	0.000
X^d_{nc}						
Same bandwidths	138.96	0.000	12.38	0.000	109.39	0.000
Different bandwidths	138.96	0.000	12.38	0.000	109.38	0.000
X^d_{hoh}						
Same bandwidths	2.32	0.005	19.40	0.000	41.02	0.000
Different bandwidths	2.16	0.005	19.50	0.000	41.24	0.000
X^d_{u5c}						
Same bandwidths	0.18	0.281	4.05	0.000	2.75	0.000
Different bandwidths	0.04	0.283	4.05	0.000	2.91	0.000
Marginal density						
Same bandwidths	913.91	0.000	809.30	0.000	-252.00	0.000
Different bandwidths	196.86	0.000	-252.99	0.000	-388.25	0.000
Joint density						
Same bandwidths	na	na	655.10	0.000	666.90	0.000
Different bandwidths	-120.46	0.000	450.43	0.000	453.00	0.000

	Binomial models sub-samples					
	$\widehat{f}_1(\cdot) = \widehat{f}_2(\cdot)$		$\widehat{f}_1(\cdot) = \widehat{f}_3(\cdot)$		$\widehat{f}_2(\cdot) = \widehat{f}_3()$	
	Test Stat	Sig.	Test Stat	Sig.	Test Stat	Sig.
U^d						
Same bandwidths	5.78	0.000	23.94	0.000	9.51	0.000
Different bandwidths	5.70	0.000	23.73	0.000	9.79	0.000
X^c_{age}						
same bandwidth	-451.29	0.000	-383.51	0.000	-266.86	0.000
Different bandwidths	120.49	0.000	81.73	0.000	100.94	0.000
X^d_{nc}						
Same bandwidths	70.14	0.000	10.68	0.000	93.39	0.000
Different bandwidths	70.14	0.000	10.68	0.000	93.39	0.000
X^d_{hoh}						
Same bandwidths	8.32	0.000	1.53	0.008	14.12	0.000
Different bandwidths	8.16	0.000	1.78	0.008	14.39	0.000
X^d_{u5c}						
Same bandwidths	3.00	0.000	21.17	0.000	11.28	0.000
Different bandwidths	2.88	0.000	21.29	0.000	11.53	0.000
Marginal density						
Same bandwidths	772.56	0.000	525.60	0.000	515.84	0.000
Different bandwidths	-153.93	0.000	-33.67	0.000	457.00	0.000
Joint density						
Same bandwidths	742.74	0.000	508.02	0.000	491.65	0.000
Different bandwidths	-163.16	0.000	349.05	0.000	349.05	0.000

[1] $\widehat{f}_1(\cdot)$ denotes the estimated kernel unconditional density for the high qualification females group. $\widehat{f}_2(\cdot)$ and $\widehat{f}_3(\cdot)$ denote the estimated kernel unconditional density for the medium and low qualification females groups respectively.

[2] The test with same bandwidth uses the pooled sample estimated kernel density, the null hypothesis assumes that unconditional density is the same in the two categories.

[3] The test with different bandwidth uses the kernel unconditional density in the sub-samples of education level. The null hypothesis is that the density is the same in the two sub-samples.

[4] na=missing test results.

2.6 Summary and conclusions

This research estimates female labour force participation decision in the UK using a sample of females in the working age 16-59 from the UK labour force survey in 2007. The objective of the research is to estimate the conditional probability of females working status, whether employed, unemployed or economically inactive, using a mixed type independent variables set that includes the basic factors that affect female propensity to work, like fertility (the number of dependent children and the number of under 5 years old children) and education level. The labour economic literature and the applied econometrics literature are rich with empirical research on the female labour force participation topic, but most of the available work uses parametric techniques. The contribution of this research is that it applies nonparametric kernel techniques for the first time to a UK individual level female labour force participation data.

The nonparametric technique that are used is kernel method with fixed bandwidth (KMWFB) utilising the mixed data types (MDT) estimation framework. The kernel method estimates are compared with the parametric multinomial logit model. The comparison uses discrete choice smoothing specification tests that are developed recently in nonparametric estimation techniques. The results are supported with graphical presentation to illustrate the differences between the sample predictions that are produced by each technique. Density equality tests are used to test whether the unconditional kernel density of the variables in the model are the same, using the categories of the education level variable in the pooled sample in addition to testing the equality of the unconditional densities among the sub-samples. The division of the sample of all females into smaller sub-samples is useful to overcome the computation problem that encounters the estimation of the kernel conditional density and in performing the smoothing kernel specification tests.

The research uses different specifications for the continuous independent variable (age) in the parametric logit model. The quadratic form in age, which strongly supported in the literature of female labour force participation, is used. In addition a semi-parametric form is used, which replaces the quadratic form in age with year of age dummy variables. The kernel conditional density estimate shows an outstanding performance in estimating female propensity to work compared with the parametric logit models that use those two forms. The correct specification smoothing tests failed to accept the parametric logit models with quadratic form and with semi-parametric form with very high significance. The sample pointwise predictions of the parametric models show that the logit predicted probabilities are consistent with the theory of female labour market input in respect to the effect of fertility and education. But they are highly restrictive and can only be interpreted in a very narrow context. The conditional probabilities that predicted by the kernel conditional density function, in contrast to the logit model, are highly consistent with the female labour force participation theory, and explain the heterogeneity in the sample better.

Based on those outcomes the research extends the semi-parametric form to a form that uses interaction terms of age dummies with number of children and eduction. The interaction terms are added into two stages; (i) adding the interaction terms of age with the number of dependent children and with education level, (ii) adding the interaction terms with the head of household dummy to the interaction terms that are used in (i). The objective of the interaction terms models is to make the parametric logit model capture the same properties that captured by the kernel estimator. However, despite that the smoothing specification tests show a decrease in the differences between the probabilities that are estimated by the kernel estimator and the interaction term model in (ii), the interaction terms model encounter a problem of multicollinearity, due to the large number of regressors that are included, which makes the model ineffective.

The kernel conditional density estimate shows outstanding performance in estimating the conditional probabilities of female propensity to work. The properties that are captured by the kernel conditional density estimate are not captured by the discrete choice parametric model with any of the specifications that used in this research. The plots of the predicted probabilities demonstrate the outstanding qualities of the kernel estimate compared with the discrete choice parametric model. The information provided by the plots is useful, where they show that the heterogeneity between the females that affect propensity to work varies across age even for females with the same number of children and qualifications. The results of the kernel estimator are highly consistent with the theory of female labour force participation in the literature.

The kernel conditional density estimates show that female propensity to work is highly dependent on the number of dependent children, particularly with under 5 years old dependent children, and varying between females with different qualifications. The effect of number of children on the propensity to work is negative, and positive in the propensity of being economically inactive. In contrast, the effect of education level is positive in the propensity to work, where females with higher qualifications have higher tendency to be employed. Females with no dependent children have always high propensity to be employed. The propensity to work of females with dependent children is lower than the propensity of females with one dependent child, particularly in the age group 25-35 where the differences are extremely high. However, for the females after the working premium age, 45-59, the propensity to work seems indifferent with respect to the number of dependent children. This may indicate that females in this age group, 45-59, have good working experience, are mature with regard to their responsibilities at home and work, so that they could participate in the labour force and work to support their dependent children. However, the propensity to work for females in the young group, 25-35, drops substantially with the number of dependent children because they are more likely to have under 5 years old dependent children.

Chapter 3

The Estimation of the Grouped Time Conditional Hazard Rate Using Kernel Method with Mixed Data Types

Chapter Abstract

This chapter suggests a new kernel hazard rate estimator for discrete time single state transition models. The estimator is developed to estimate the hazard rate conditioning on mixed data types time invariant independent variables. Most of the duration models that are estimated in economics are measured in discrete (grouped) time unit and estimated conditional on continuous and categorical variables together. The research develops an estimator that can be the kernel nonparametric counterpart of the cloglog discrete time proportional hazard rate estimator of Jenkins (1995).

The estimator is extended from the continuous time external kernel hazard estimation framework of Watson and Leadbetter (1964a,b), and uses the sub-survival kernel function estimation method for the conditional hazard of Beran (1981). The discrete time hazard estimator of Tutz and Pritscher (1996), and the mixed independent variables smoothing technique for the conditional distribution function of Li and Racine (2008), provide good resources to develop the discrete time kernel hazard estimators successfully. The estimation of the kernel hazard rate in discrete time has many advantages over the continuous time estimator, where the Boundary Bias problem is treated better and reduced.

A Monte Carlo simulation study is constructed to examine the new kernel hazard estimator. The naive bandwidths are used in the simulation to reduce the computation cost. The new estimator shows good performance in spite that the naive bandwidths reduces the quality of kernel estimates in the simulation study compared with the PH models. A data-driven method to estimate the bandwidths of the discrete time kernel hazard is needed to improve the estimator. Extending the estimator to include time varying covariates is left for future research.

Key Words: Nonparametric, hazard rate, survival function, grouped time hazard, Weibull hazard, proportional hazard, external kernel hazard, Beran kernel hazard, single state transition, , sub-survival function, naïve bandwidths.

JEL Classification: C14, C15, C24, C25, C51, C63.

3.1 Introduction

The estimation of duration models, or survival models, using kernel method is one of the sophisticated, but also the less developed, areas of nonparametric kernel estimation method in econometrics. The hazard rate and the survival function have many applications in economics. For example, in modelling the duration of the states for individuals in given labour market states - unemployment, employment, out of the labour force etc, in addition to applications in trade, finance, poverty and many other fields. In contrast to other disciplines, because of the nature of the data being analysed, survival models in economics and social sciences are often estimated in discrete time units, i.e. in intervals of weeks, months, quarters, ...etc. Although the underlying data generation process in the population is in a continuous time, the hazard can only be estimated in the discrete (grouped) time. The discrete time hazard estimator is then known also as the **interval censored data hazard estimator**. The hazard is estimated using the observed number of transitions in each interval without considering the sequence of the transitions in the underlying continuous time inside the intervals.

The development of the duration models in the discrete time must not ignore the underlying continuous time structure of the process. The discrete time model can be developed using the information from the observed part of the continuous time, which are the time index points (the points at which the observation are taken). The grouped time hazard estimator is theoretically not affected by whether the time index points are defined at the beginnings or at the ends of the intervals. The continuous time survival function at each time index point equals the discrete time survival function at the interval corresponding to that point. This explains the use of the continuous time survival function to construct the hazard rate and the survival function formulas in the grouped time transition model.

The estimation of the discrete time hazard rate is well defined in the literature of the fully parametric and semi-parametric estimation techniques in econometric. Standard references are Lancaster (1990), Van den Berg (2001), Wooldridge (2002), Cameron and Trivedi (2005) and Jenkins (2005). But all of them say very little about the methods to estimate a hazard rate using the kernel method, which are known as the **kernel hazard (KH) estimators**. The interest in the nonparametric kernel method is growing because it is showing impressive performance in many applications, particularly in the models that include discrete variables. Chapter 2 shows the high quality results that are possible using the kernel method estimators. The use of the kernel method in Survival Analysis is quite popular in medicine and industry, where well established estimators for the conditional and the unconditional hazard and survival functions are available. However, most of those estimators are suitable for the continuous time hazard models and cannot be applied directly to the discrete time hazard cases. So, to apply kernel method to duration models in economics new estimators are need to be developed.

There are two ways to estimate a continuous time hazard rate using the kernel method, the **external approach** and the **internal approach**, see Nielsen and Linton (1995) and Spierdijk (2008). The external approach tries to estimate the hazard rate directly by smoothing the data, while the internal approach attempts to estimate the cumulative hazard function first. Intuitively, those two approaches are equivalent to the kernel method approaches for density and cumulative distribution functions estimation. In Chapter 1 it is shown that the estimation of the density function is easier than the estimation of the cumulative distribution function. A considerable attention must be paid to the cumulative distribution function estimator to avoid producing estimates that lay outside the $[0, 1]$ range. For discrete variables an estimator of a cumulative distribution function is little important, where the discrete variable density estimator is sufficient. The same is true in the hazard estimation. Where, we argue that developing an estimator for the discrete time hazard is sufficient and there is no need to develop an estimator for the discrete time cumulative hazard. The external approach, suggested by Watson and Leadbetter (1964a,b), is easier to be adopted in the discrete time because it can be extended to this time setting easily and directly. Accordingly, we will focus on the external approach of kernel hazard estimation only to develop the discrete time estimator.

Using the kernel method to estimated the hazard in discrete time appeared to be rarely considered in the literature, we find only Tutz and Pritscher (1996), who suggest a kernel hazard estimator for the Life Tables discrete data. Since the development of the kernel smoothing method with mixed data types by Li and Racine (2003), Hall et al. (2004), Hall et al. (2004), Li and Racine (2008), and Li and Racine (2007), a discrete time hazard with mixed independent variables appeared to be not attempted.

The objective of this research is to introduce a kernel estimator for the discrete time single state transition conditional hazard rate model, conditioning on mixed type independent variables. The suggested estimator extends the existing external kernel hazard estimator of the continuous time transition models. The construction of the estimators in the continuous time and the grouped time setting is studied to show the relationship between the density, distribution, survival and hazard functions in these two time settings clearly. The grouping scheme of the continuous time is studied to demonstrate our choice of the kernel function that used to smooth the grouped time variable in our estimator. The research deals with the right censoring problem only, problems like left truncation, left censoring and right truncation are not considered. The technique is developed when only time invariant independent variables are included. The developed kernel estimator is compared with the popular discrete time proportional hazard method of Jenkins (1995) using the simulation technique. The asymptotic properties of the kernel hazard are not discussed and left for future research.

The research considers the duration models in the homogeneous case only, i.e when it is assumed that there is no effect of the unobserved heterogeneity in the model. The unobserved heterogeneity problem is widely considered in the duration models in economics, the mentioned references above to Horowitz (1999), Baker and Melino (2000), Abbring and Van Den Berg (2007) discuss this topic thoroughly in the parametric hazard rate estimation framework. But not discussed sufficiently in kernel estimation techiques. To control for the effects on the unobserved components, v, in the parametric PH estimation framework, a function that depends only on the unobserved components, $g(v)$, is introduced in the model and a distribution assumption that regards, v, is added. Abbring and Van Den Berg (2007) shows that the asymptotic distribution of the unobserved heterogeneity is Gamma but in the literature there are models that use Gaussian or nonparametric distribution assumptions. The unobserved components are assumed to be uncorrelated with the independent variables and time fixed. In the parametric estimation of the duration model v is treated as random effects component. The parametric estimation framework controls for the effect of the unobserved heterogeneity by using techniques that developed by the mixing the distribution assumptions of the transition time, which is discussed more in the next section and the next chapter, and the distribution assumption of the unobserved heterogeneity (a mixture models estimation technique).

In the kernel estimation framework, the literature is little informative about the ways to handle the unobserved heterogeneity problem. We find only a claim by Froelich (2006), in a comparison between the nonparametric regression and the binary logit model, says that the kernel regression is less affected by the problem of the unobserved heterogeneity than the parametric discrete choice models. After a sufficient search in this subject and in the ways of estimating mixed models using the kernel method, we find that it is very difficult theoretically to develop the discrete time hazard estimator on a mixed distribution assumptions. In addition to that, smoothing random effects is shyly discussed in kernel estimation framework. However, we distinguish in the literature few techniques that are promising into providing solutions to the mixed distributions issue, like the re-weighted kernel density method of Hall and Turlach (1999) and Hazelton and Turlach (2007) which is discussed more in Wang and Wang (2007), and the constrained kernel density technique of Chen (1997) and Hall and Presnell (1999), which based on the empirical likelihood method, with a similar technique in the kernel hazard rate estimation presented by Hall et al. (2001). All those estimation methods are discussed in Chapter 2. For the conditional hazard rate, the weighted local constant estimator of Cai (2002), see Li and Racine (2007), and the constrained nonparametric kernel regression method of Racine et al. (2009) are all attractive estimators that could be considered. Those are a few kernel methods that we think they may be useful to improve the conditional kernel hazard estimator that suggested in this research.

With the short research that available in this field, we consider this research one of the very early attempts to drive a kernel estimator for discrete time hazard rate using the mixed data types (MDT) estimation framework. The main contribution of this research is that it suggests and examines a new estimator in the well known, novel, kernel estimation technique. We expect to find many applications to our new estimator in economics and other fields. The chapter is organized as follow: Section 3.2 interprets the main terms in the Survival Analysis literature and the Economic Duration models, Section 3.3 and 3.4 illustrate the continuous time kernel hazard estimators and our discrete time kernel hazard estimator respectively. Then a simulation study to compare the kernel discrete time hazard with the discrete time PH models is designed in section 3.5. The last section summarises the conclusions of the research.

3.2 Definitions

The duration models as defined in Cameron and Trivedi (2005, page 573) are the **"models of the length of the time spent in a given state before transition to another state"**. This type of economic problems can be addressed in many forms, for example: in the individual level data, the duration of employment and unemployment spells; in the household level data, modelling the length of time that a household stay under the poverty line, and so on. The period that a unit of research - we refer to them as the individuals - spend in a particular state, is known as the **spell length** or the **duration length**. A **transition** means that a successful shift to another state is observed for an individual in the sample. The period that the individuals spend in the state and the probability of transition conditional of the period spent in the state are of special interest in the duration models. The relationship between the probability of transition and the spell length is known as the duration dependence, it can be negative if the probability decreases with the increase in the duration of the spell, positive if the opposite or, rarely, they can be not interdependent. The probability of staying in the state after a certain period is known as the survival probability and is measured by the survival function, which is the complement of the cumulative distribution function. The survival function will play the key rule in the development of the grouped time hazard model.

In the duration models, we are specially interested in the hazard rate, Cameron and Trivedi (2005, page 576) defines the hazard as **"the instantaneous probability of leaving a state conditional on survival time"**. The hazard is the density function divided by the survival function. The estimation of the hazard rate, and also the survival function, requires assumptions about the distribution of the elapsed time in the spell. The assumptions focus on the distribution function of the continuous time. The estimation of the process in the grouped time though, must consider those distribution assumptions as assumptions of the underlying continuous time. To estimate a duration model in the continuous time, the time variable can assumed to have Exponential, Weibull, Gompertz, Log-normal, Log-logistic, or Gamma distribution. In the grouped time models there are less options, where only Exponential, Weibull or Log-logistic assumptions are used for the underlying continuous time, because those are the distribution assumptions that work with the **proportional hazard (PH) estimation framework**. The less number of assumptions in the grouped time case is because only the proportional hazard estimation framework is used in discrete time transition models. Whilts in the continuous time duration models both the **accelerated failure time (AFT) estimation framework** and the proportional hazard estimation framework are used, see Jenkins (1995, 2005), Sueyoshi (1995) and Allison (1982).

The proportional hazard estimation framework assumes that the conditional hazard is a product of two functions: i) a baseline hazard that depends on time only, this function measures the duration dependence of the spell. Although it can be identified via the distribution assumptions, the model can still be estimated if it is not identified, where a Partial Likelihood, a Marginal Likelihood or a Profile Likelihood method can be used. ii) the observed heterogeneity function, which is a function on the contribution of the independent variables to the baseline hazard, this function is on the independent variables only and independent of time. The AFT on the other hand, estimates the conditional hazard in a regression form. Therefore, all the terms in the AFT model, like the intercept and the coefficients, should be identified. In the continuous time an estimated conditional hazard under one of the distribution assumption above in a PH form can easily be compared with its corresponding AFT form, the coefficients of the independent variables can easily be re-parametrize from one form to another.

The grouping of the time affect the estimation of the baseline function only in the PH model. The contribution of the independent variables to the baseline hazard is unchangeable with the grouping of the transition time variable. The observed heterogeneity function is unaffected by the grouping because it is independent of time. So, the coefficients of the grouped time PH model are the same as the coefficients of the continuous time PH and can be re-parametrized to the AFT form. The grouped time models have more advantages over the continuous time models. Where the occurrence of tie transitions in the sample do not complicate the estimation. If there is any lake of identification problem from the grouping it affects the baseline hazard function only, but is treated in the grouped time PH model by introducing a semi-parametric or a non-parametric specification of the baseline hazard function in the model. Those facts are highly considered in the development of the simulation study in the last section of this research.

The censoring problem is the most common problem that affects the estimation of the duration models, that occurs when the exact start date of the spell or the date of transition is not observed. Generally, the censoring occurs when the spell is not observed in full and the information about the length of time in the state is incomplete for some units in the sample. In other words, the information is unavailable for an individual after a certain period in the state, so the transition is not observed for that individual. The right censoring is problematic in the duration models and uninformative, Klein and Moeschberger (2003), and complicates the estimation because all that known about the right censored individuals is that the transition will occur to them at some point in the future.

There are other forms of the censoring problem, where the spell can be censored from the left if the information about the beginning date of the spell is not available, or interval censored if the transition is known to occur inside intervals. The left censoring is more difficult to handle than the right censoring, it causes more serious bias in the model. Generally it is preferred to drop the left censored observations from the sample. The interval censoring on the other hand, can be removed by grouping the time, where the grouped time hazard model is also known as the interval censored hazard model, Mátyás and Sevestre (2008). So this is an additional advantage for the grouping of the duration time, in addition to the advantage of handling of the tie observations problem that complicats the estimation of continuous time duration models.

3.3 Continuous Time Kernel Hazard Estimation

Let Y_i^c denotes the continuous transition time for individual i, $i = 1, 2, 3, ..., n$. Y_i^c is an iid nonnegative continuous random variable with a density function f and a distribution function F. To handle the right censoring problem in the sample, define an iid random right censoring continuous time variable C_i^c with a density function r and a distribution function R. The transition of individual i is observed if $Y_i^c \leq C_i^c$, otherwise the spell is right censored. The observed time for each individual in the sample is $T_i^c = \min(Y_i^c, C_i^c)$, with a density g and a distribution function G, but both depend on the density and the distribution functions of Y_i^c and C_i^c, where $1 - G = (1 - F)(1 - R)$. The uncensored observations are defined in the sample by an indicator function $\triangle_i = I(Y_i^c \leq C_i^c)$. This is defined as the **competing risks right censoring process**, which is commonly occurs in economic duration models.

The functions of Y_i^c are of the special interest in the duration model, since they are the functions of the true transition variable, but the process is not completely observed due to the right censoring problem in the data. In the duration analysis the sample consists of $\{T_i^c, \triangle_i, \mathbf{X}_i\}_{i=1}^n$, where $\mathbf{X}_i$ is a vector of mixed type independent variables. The model can be estimated parametrically or nonparametrically. The kernel method can be used, where the kernel hazard rate estimators have means to correct the right censoring problem. There are many kernel estimators for the hazard and the survival functions in the continuous time, but we will illustrate the estimators that developed from the external kernel hazard only in the following subsections[1].

[1]The notation that used in this section is related to the discussion in the section only, we use different notation for the continuous time transition variable that is being grouped in the following sections.

3.3.1 The Unconditional Continuous Time Kernel Hazard

The unconditional hazard rate in continuous time is defined as:

$$\phi\left(t^{c}\right)=\frac{f\left(t^{c}\right)}{1-F\left(t^{c}\right)}=\Lambda^{'}\left(t^{c}\right), \tag{3.1}$$

where $\Lambda\left(t^{c}\right)=-\log\left(1-F\left(t^{c}\right)\right)$ is the cumulative hazard function and $\Lambda^{'}\left(t^{c}\right)$ is its derivative. The hazard rate is estimated parametrically by assuming that $f\left(t^{c}\right)$ follows Exponential, Weibull, Gompertz, Log-normal, Log-logistic, or Gamma distribution as illustrated above, or it could be estimated nonparametrically using the sample frequencies or the kernel method.

Among the estimators that are discussed by Rice and Rosenblatt (1976) there are two kernel estimators of the hazard rate for continuous time transition models that are relevant to our discussion in this research. The first estimator is the ratio between the estimated kernel density and the estimated kernel survival function as follows, see Rice and Rosenblatt (1976, page 62, Eq (2.9) and (2.10)):

$$\widehat{\phi}_{1}\left(t^{c}\right)=\frac{\widehat{f}_{h}\left(t^{c}\right)}{1-\widehat{F}_{h}\left(t^{c}\right)}. \tag{3.2}$$

The second estimator is based on Watson and Leadbetter (1964a) direct estimator of the hazard rate and takes the form:

$$\phi_{2}\left(t^{c}\right)=\int h_{0}^{-1}k\left(\frac{s-t^{c}}{h_{0}}\right)\frac{dF_{n}\left(s\right)}{1-F_{n}\left(s\right)}, \tag{3.3}$$

where F_n is a modified sample distribution function. The modification is necessary to avoid dividing by 0 in the hazard estimator in Eq (3.3). The sample distribution function uses the order statistics of the duration time, since the sequence of the transitions of the individuals in the sample is as important as the number of the transitions. In other words, in the continuous time hazard estimation, we sort the individuals in the sample so that the earliest transition be first followed by the individual with the second earliest transition and so on. The transition time is sorted ascendantly from the lowest to the highest. The order statistics of transition time variable, $T^{c}_{(i)}$ where $T_{(1)}\leq T_{(2)}\leq T_{(3)}\leq\cdots\leq T_{(n)}$ is used in the kernel estimator instead of T_i.

The sample distribution function is modified for the order statistics $T_{(i)}$ to avoid having $F_{(n)}=1$ since this produces a sample survival function $1-F_{(n)}=0$ which produces unspecified value in the hazard estimator. The sample distribution function as shown by Watson and Leadbetter (1964a) could be modified to $F_{n}\left(T^{c}_{(i)}\right)=\frac{i}{n+1}$ without affecting the precision of the kernel hazard estimator, just by multiplying the usual empirical sample distribution function by $\frac{n}{n+1}$.

The kernel function, $k\left(\cdot\right)$, in Equations (**??**) and (3.3) is a continuous variable kernel function that locally weight the relative difference of the order statistics $T_{(i)}$ from t^{c}

relitively to h_0. The scalar h_0 is known as the smoothing parameter or the bandwidth. Kernel functions for continuous variables have different forms and orders, see Wand and Jones (1995). The kernel hazard estimators are designed to be less sensitive to the form of the kernel function, and use low order kernel function, usually not exceeding the second order. The higher order kernel functions are difficult to use due to the high computation cost of them. However, the method to estimate the bandwidth, h_0, are carefully considered because the bandwidth value affects the variance and the bias of the predictions directly.

Let $z = \frac{T^c_{(i)} - t^c}{h_0}$ be the distance between the ordered sample observation $T^c_{(i)}$ and the value t^c proportional to h_0. The kernel function should satisfy the following consistency conditions:

1. $\int k(z)\, dz = 1$, and $\int z k(z)\, dz = 0$.
2. Symmetry: $k(z) = k(-z)$.
3. $\int z^r k(z)\, dz = \tau_r \neq 0$. for $r > 1$.
4. As $n \longrightarrow \infty$, $h_0 \longrightarrow 0$ and $nh_0 \longrightarrow \infty$.

Watson and Leadbetter (1964b) show that the continuous variable kernel functions under this properties are legitimate to the hazard rate estimator.

The hazard estimator smooths the order statistic of the elapsed time in the sample $T_{(i)}$ and uses the empirical sample distribution function as follows,

$$\widehat{\phi}_2(t^c) = \sum_{i=1}^{n} \widehat{h}_0^{-1} k\left(\frac{T_{(i)} - t^c}{\widehat{h}_0}\right) \frac{d\widehat{F}_n\left(T_{(i)}\right)}{1 - \widehat{F}_n\left(T_{(i)}\right)}, \tag{3.4}$$

$$= \sum_{i=1}^{n} \widehat{h}_0^{-1} k\left(\frac{T_{(i)} - t^c}{\widehat{h}_0}\right) \frac{1/(n+1)}{1 - i/(n+1)}, \tag{3.5}$$

$$= \sum_{i=1}^{n} \widehat{h}_0^{-1} k\left(\frac{T_{(i)} - t^c}{\widehat{h}_0}\right) \frac{1}{n-i+1}, \tag{3.6}$$

The bandwidths $\widehat{h}_0$ is estimated using cross-validation method. Few research, however, use the Plug-in bandwidth of the density function when the computation of the CV bandwidth is difficult. The quantity $n-i+1$ at each sample ordered transition time $T_{(i)}$ is the number of individuals **at risk** at time $T_{(i)}$. This estimator is know as the **continuous time external kernel hazard** in the continuous time kernel hazard estimation literature, as it is a direct kernel estimator of the hazard in contrast to $\widehat{\phi}_1(t^c)$. Where $\widehat{\phi}_1(t^c)$ estimates the kernel density and the kernel distribution functions first before estimating the hazard rate.

For right censored observations the formula can be corrected by adding the right censoring indicator, so that the continuous time external kernel hazard estimator is modified as

$$\widehat{\phi}_2\left(t^c\right)=\widehat{h}_0^{-1}\sum_{i=1}^{n}k\left(\frac{T_{(i)}^c-t^c}{\widehat{h}_0}\right)\frac{\Delta_{(i)}}{n-i+1}. \tag{3.7}$$

Watson and Leadbetter (1964b) and Rice and Rosenblatt (1976) show that the bias of $\widehat{\phi}_1\left(t^c\right)$ and $\widehat{\phi}_2\left(t^c\right)$ are slightly different and take the forms:

$$Bias\left(\widehat{\phi}_1\left(t^c\right)\right)=\frac{f''\left(t^c\right)}{1-F\left(t^c\right)}\frac{h_0^2}{2}\int u^2k\left(\frac{u-t^c}{h_0}\right)du+o\left(h_0^2\right), \tag{3.8}$$

and

$$\begin{aligned}Bias\left(\widehat{\phi}_2\left(t^c\right)\right) &= \phi''\left(t^c\right)\frac{h_0^2}{2}\int u^2k\left(\frac{u-t^c}{h_0}\right)du+o\left(h_0^2\right), \\ &= Bias\left(\widehat{\phi}_1\left(t^c\right)\right)+\widehat{\Phi}\left(t^c\right).\end{aligned} \tag{3.9}$$

where $\phi''\left(t^c\right)$ is the second derivative of the true hazard, f'' is the second derivative of the true density of the transition time, F is the true distribution function and

$$\widehat{\Phi}\left(t^c\right)=\left(3\phi\left(t^c\right)\phi'\left(t^c\right)+\phi\left(t^c\right)^3\right)\left[\frac{h_0^2}{2}\int u^2k_h\left(u\right)du+o\left(h_0^2\right)\right].$$ [2]

The variances $\widehat{\phi}_1\left(t^c\right)$ and $\widehat{\phi}_2\left(t^c\right)$ are the same and take the form:

$$Var\left(\widehat{\phi}_1\left(t^c\right)\right) = Var\left(\widehat{\phi}_2\left(t^c\right)\right) \tag{3.10}$$

$$= \frac{1}{nh_0}\frac{\phi\left(t^c\right)}{1-F\left(t^c\right)}\int k^2\left(\frac{u-t^c}{h_0}\right)du+o\left(\frac{1}{nh_0}\right). \tag{3.11}$$

3.3.2 The Conditional Continuous Time Kernel Hazard Estimator

An estimator of the kernel hazard conditioning on $\mathbf{x}$, where the vector $\mathbf{x}$ includes two types of variables, continuous independent variables $\mathbf{x}^c$ and discrete independent variables $\mathbf{x}^d$, is introduced by Beran (1981). The estimator uses a so-called **sub-survival function**, which is the survival function of the individuals in the risk sets just after each observed failure time in the sample. The conditional hazard rate is estimated by taking the ratio of the conditional probability of transition and the conditional sub-survival probability for the individuals in the risk set. The estimator is known as **Beran kernel hazard (BKH) estimator** and has the following form, see Beran (1981, page 8):

$$\widehat{\phi}^{BKH}\left(t^c\left|\mathbf{x}\right.\right)=\frac{\widehat{S}_{\mathbf{h},1}\left(t^c\left|\mathbf{x}\right.\right)-\widehat{S}_{\mathbf{h},1}\left(t^{c+}\left|\mathbf{x}\right.\right)}{\widehat{S}_{\mathbf{h},2}\left(t^c\left|\mathbf{x}\right.\right)}, \tag{3.12}$$

where $\widehat{S}_{\mathbf{h},1}\left(t^c\left|\mathbf{x}\right.\right)$ is the survival probability exactly at t^c conditioning $\mathbf{x}$ and $\widehat{S}_{\mathbf{h},1}\left(t^{c+}\left|\mathbf{x}\right.\right)$ is survival probability just after t^c conditioning on the same $\mathbf{x}$. The survival function

[2]Combined by the writer, see Rice and Rosenblatt (1976).

is a non-negative monotonic non-increasing function in the range $[0, 1]$, which means that the terms in the nominator is less than or equal 1. Because of the one-to-one relationship between the survival and the distribution functions, the subtraction of the conditional survival functions, $\widehat{S}_{\mathbf{h},1}$, in the nominator approximates the density of the transition variable. The subscript 1 in the sub-survival function estimator denotes that the sub-survival function $\widehat{S}_{\mathbf{h},1}$ is estimated for the uncensored individuals only.

The sub-survival function in the denominator, $\widehat{S}_{\mathbf{h},2}\left(t^c\,|\mathbf{x}\right)$, is the conditional survival probability of the individuals in the risk set at t^c, some individuals in this set could be censored in later times (after t^c) but this is considered only when the hazard is estimated at those times.

The sub-survival functions are estimated by nonparametric kernel regression techniques. Beran (1981) uses the Nadaraya-Watson regression for the individuals in the risk set, and indicator functions are used to distinguish the observations in the risk set at each $T_{(i)}$ in the sample. The sub-survival functions are given by the formulae

$$\widehat{S}_{\mathbf{h},1}\left(t^c\,|\mathbf{x}\right) = \sum_{i=1}^{n} \frac{W(\mathbf{x}, \mathbf{X}_{(i)}, \widehat{\mathbf{h}}) I\left(T^c_{(i)} \geq t^c; \triangle_{(i)}= 1\right)}{\sum_{i=1}^{n} W(\mathbf{x}, \mathbf{X}_{(i)}, \widehat{\mathbf{h}})}, \tag{3.13}$$

and

$$\widehat{S}_{\mathbf{h},2}\left(t^c\,|\mathbf{x}\right) = \sum_{i=1}^{n} \frac{W(\mathbf{x}, \mathbf{X}_{(i)}, \widehat{\mathbf{h}}) I\left(T^c_{(i)} \geq t^c\right)}{\sum_{i=1}^{n} W(\mathbf{x}, \mathbf{X}_{(i)}, \widehat{\mathbf{h}})}, \tag{3.14}$$

respectively

The same estimator can be written in term of the sub-distribution functions as follows

$$\widehat{\theta}^{BKH}\left(t^c\,|x\right) = \frac{\widehat{F}_{\mathbf{h},1}\left(t^{c+}\,|\mathbf{x}\right) - \widehat{F}_{n,1}\left(t^c\,|\mathbf{x}\right)}{1 - \widehat{F}_{\mathbf{h},2}\left(t^c\,|\mathbf{x}\right)}, \tag{3.15}$$

where

$$\widehat{F}_{\mathbf{h},1}\left(t^c\right) = \sum_{i=1}^{n} \frac{W(X_{(i)}, \widehat{\mathbf{h}}) I\left(T^c_{(i)} \leq t^c; \triangle_{(i)}= 1\right)}{\sum_{i=1}^{n} W(X_{(i)}, \widehat{\mathbf{h}})} \tag{3.16}$$

$$\widehat{F}_{\mathbf{h},2}\left(t^c\right) = \sum_{i=1}^{n} \frac{W(\mathbf{x}, \mathbf{X}_{(i)}, \widehat{\mathbf{h}}) I\left(T^c_{(i)} \leq t^c\right)}{\sum_{i=1}^{n} W(\mathbf{x}, \mathbf{X}_{(i)}, \widehat{\mathbf{h}})}. \tag{3.17}$$

The function $W(\mathbf{x}, \mathbf{X}_{(i)}, \mathbf{h})$ is a product of the kernel functions of the mixed variables in $\mathbf{x}$. $\mathbf{h}$ is a vector of the bandwidths of the independent variables only with length $q + p$. The vector $\mathbf{h}$ includes a sub-vector of the bandwidths of the continuous independent variables $\mathbf{h}_{\mathbf{x}^c} = [h_1, h_2, ..., h_q]$ and a sub-vector of the bandwidths of the discrete independent variables $\gamma_{\mathbf{x}^d} = [\gamma_1, \gamma_2, ..., \gamma_p]$. The product of the kernel functions multiply the univariate kernel functions of the continuous independent variables

$h_s^{-1} w\left(\frac{X^c_{(i)s} - x^c_s}{h_s}\right)$, $s = 1, 2, .., q$, and the univariate kernel function of the discrete independent variable $l\left(X^d_{(i)r}, x^d_r, \gamma_r\right)$, $r = 1, 2, ..., p$, where the sample observations of each variable are smoothed around their corresponding value in the vector $\mathbf{x}$. So, $W(\mathbf{x}, \mathbf{X}_{(i)}, \mathbf{h})$ has the following form,

$$W(\mathbf{x}, \mathbf{X}_{(i)}, \mathbf{h}) = \prod_{s=1}^{q} h_s^{-1} w\left(\frac{X^c_{(i)s} - x^c_s}{h_s}\right) \prod_{r=1}^{p} l\left(X^d_{(i)r}, x^d_r, \gamma_r\right),$$

note that $\mathbf{x} = \left[x^c_1, \cdots x^c_q, x^d_1, \cdots x^d_p\right]$.

The sample is sorted by the order statistic $T_{(i)}$ only, so $\mathbf{X}_{(1)}$ is the vector that corresponds to $T_{(1)}$ but not necessary be the vector that include the least values of $\mathbf{X}$ in the sample. The Nadaraya-Watson weights for the sub-survival functions are

$$\mathbf{A}(\mathbf{x}, \mathbf{X}_{(i)}, \mathbf{h}) = \frac{n^{-1} W(\mathbf{x}, \mathbf{X}_{(i)}, \mathbf{h})}{\sum_{i=1}^{n} n^{-1} W(\mathbf{x}, \mathbf{X}_{(i)}, \mathbf{h})}.$$

Beran (1981) uses this hazard estimator to predict the survival function as follows:

$$\widehat{S}\left(t^c \,|\mathbf{x}\right) = \prod_{T_{(i)} \leq t} \left(1 - \widehat{\phi}^{BKH}\left(T^c_{(i)}, |\mathbf{x}\right)\right)^{\triangle_{(i)}}. \tag{3.18}$$

The second conditional kernel hazard estimator uses Nadaraya-Watson weights also, but it can be considered as a direct extended version of the unconditional external kernel hazard that presented above. The sub-survival functions in the formula in Eq (3.12) are integrated in the unconditional kernel hazard formula in Eq (3.7) to give a hazard estimator as follows:

$$\widehat{\phi}^{CEKH}\left(t^c \,|\mathbf{x}\right) = \widehat{h}_0^{-1} \sum_{i=1}^{n} k\left(\frac{T^c_{(i)} - t^c}{\widehat{h}_0}\right) \frac{\mathbf{A}(\mathbf{x}, \mathbf{X}_{(i)}, \widehat{\mathbf{h}})\, \triangle_{(i)}}{1 - \sum_{j=1}^{i-1} \mathbf{A}(\mathbf{x}, \mathbf{X}_{(i)} \widehat{\mathbf{h}})}. \tag{3.19}$$

This estimator is called the **continuous time conditional external kernel hazard estimator** and denoted by $\widehat{\phi}^{CEKH}$. The same estimator can be written in a sub-distribution function form as well. It is easy to show that the unconditional external hazard estimator in Eq (3.19) follows when the Nadaraya-Watson weights are replaced by the empirical distribution function weights, i.e when we put $\mathbf{A}(\mathbf{x}, \mathbf{X}_{(i)}, \mathbf{h}) = n^{-1}$, and then

$$h_0^{-1} \sum_{i=1}^{n} k\left(\frac{T^c_{(i)} - t^c}{h_0}\right) \frac{n^{-1}\, \triangle_{(i)}}{1 - \sum_{j=1}^{i-1} n^{-1}} = h_0^{-1} \sum_{i=1}^{n} k\left(\frac{T^c_{(i)} - t^c}{h_0}\right) \frac{\triangle_{(i)}}{n - i + 1} = \widehat{\phi}^{EKH}\left(t^c\right). \tag{3.20}$$

Beran (1981) kernel conditional hazard estimator is used in Li and Doss (1995), Akritas (2004), Läuter and Liero (2006), Guilloux (2011). The conditional external kernel hazard estimator on the other hand, has received attention in few literature papers. Van Keilegom and Veraverbeke (2001) discuss $\widehat{\phi}^{CEKH}(t^c|\mathbf{x})$ with Gasser-Müller local weights instead of the Nadaraya-Watson weights, and assumes continuous partial derivatives of the sub-distribution functions with respect to x^c and t^c to illustrate the consistency conditions of the estimator. Spierdijk (2008) on the other hand, extends the external hazard estimator by replacing the Nadaraya-Watson weights, also known as local constant regression weights, in the sub-distribution functions by local linear regression weights. The local linear weights are Taylor expansion of the local constant weights. Cao and López-de Ullibarri (2007) suggests a product-type KH estimator by modifying the Nelson-Aalen estimator of the survival function. The use of Beran kernel hazard estimator and the external kernel hazard estimator, or any of their modified versions, to estimate a grouped time hazard has not come to our knowledge. We haven't been able to find any work that is relevant to any of those estimators in the discrete time duration models estimation. Accordingly, we will focus on the basic estimators only, the estimators that use the Nadaraya-Watson weights, and leave the other estimators for future research in this field.

The continuous time kernel hazard estimators have in general a bias problem, most of the published work in the context of the continuous kernel hazard estimator attempt to correct the estimator from this bias problem rather than enhancing it. The bias problem is generated in the lower bound and known as the **Boundary Bias**. The problem occurs because the kernel function allocates local weights to each observation $T^c_{(i)}$ and its nearby observations in the interval $\left(T^c_{(i)} - h, T^c_{(i)} + h\right]$. But on the lower limit the weights are given to the observations in the range $(0 - h, 0 + h]$. This shows that the continuous time hazard enforces negative values in duration variable in the lowest window, when the variable should be strictly positive.

The consequence of the Boundary Bias is that the kernel hazard estimator underestimates the hazard at the lower bound (the beginning of the spell) by almost the half, see Bagkavos (2003). There are many methods that are suggested in the literature to remove the Boundary Bias in the external kernel hazard estimator, but most of them actually reduce the bias only rather than removing it. Where in addition of extending the Nadarya-Watson weights to a local linear weights, there are some methods suggest to reduce the bias by using kernel functions with variable bandwidth, i.e. use bandwidth of the form h_{t^c} that change with the value t^c and depends on $T_{(i)}$, see Li and Racine (2007, Chapter 14). However, the variable bandwidth kernel functions are used for continuous variables only, where they are not used in the discrete or mixed variables models. Higher order kernel functions like the functions in Jones (1995) are also suggested, where they have less bias than the lower order kernel functions. But the higher order kernel functions increase the computation cost and make the estimation more difficult and less attractive.

The Boundary Bias problem is one of the reasons that make the application of the kernel hazard estimator in the continuous time transition models complicated and not attractive. But the continuous time estimator can still be extended to the discrete time hazard case. In the grouped time models the type of the kernel function that used to smooth the grouped time variable does not smooth negative values at the lower bound, so there is no source of the Boundary Bias as will be illustrated in the next section.

Generally in duration models, we gain advantages from the grouping the duration time in many aspects, like in the interval censoring problem and the ties problem as show by Hess and Persson (2009). Which are the reasons that motivated the parametric estimation of the grouped time hazard rate. In the kernel estimation the grouping remove the Boundary Bias that is associated with the continuous time estimators. This, in addition to the interval censoring and tie observations problem motivates the development of the grouped time kernel hazard estimator. However, the proportional hazard estimation framework in the continuous time models imposes very restrictive assumptions about the distribution of Y^c in the baseline hazard, and only weakly deals with the time varying covariates. Whilst the grouped time PH framework uses flexible specification of the baseline hazard and easily handle time varying independent variables, as shown by Jenkins (1995). The kernel estimator of the discrete time hazard that developed in this research is just able to include time invariant independent variables.

3.3.3 The Estimation of the Bandwidths of the Continuous Time Kernel Hazard

The least squares cross-validation (LSCV) method is proposed to estimate the bandwidths of the kernel hazard in Equations (3.7), (3.12) and (3.19). The bandwidths are estimated by minimizing the weighted integrated square error (ISE) risk function which has the form $\int \left(\widehat{\phi}\left(t^c\right) - \phi\left(t^c\right)\right)^2 \omega\left(t^c\right) dt$, where $\omega\left(t^c\right)$ is the weighting function. The risk function of the least squares cross validation of the unconditional hazard is:

$$LSCV(\widehat{h}_0) = \int \phi^2\left(t^c;\widehat{h}_0\right) dt - \frac{2}{n}\sum_{i=1}^{n}\frac{\phi^{-i}\left(t^c;\widehat{h}_0\right)}{n-i+1}\omega\left(t\right). \tag{3.21}$$

This method finds the optimum bandwidth by making a tradeoff between the bias and the variance in the model.

For the conditional external hazard Spierdijk (2007, page 2434) suggest the **integrated asymptotic mean squared error (IAMSE)** method to derive a cross validation risk function from the formula:

$$IMSE(\widehat{h}_0,\mathbf{h}) = \int\int MSE\left(t^c,\mathbf{x},\widehat{h}_0,\mathbf{h}\right) f\left(t^c\left|\mathbf{x}\right.\right) f\left(\mathbf{x}\right).dt^c.dx. \tag{3.22}$$

The above objective functions of the kernel hazard estimator is difficult to compute numerically especially for the conditional hazard estimator. The sub-survival functions in the conditional hazard formulas of $\widehat{\phi}^{BKH}$ and $\widehat{\phi}^{CEKH}$ slow the computation substantially. Alternatively, one can use the CV bandwidths of the kernel density function, or the kernel conditional density in the case of the conditional hazard, as naïve bandwidths. The bandwidths that driven from a maximum likelihood function of the conditional hazard rate can also be used. The leave-one-out method can be used in the objective function and the method is called the **maximum likelihood cross validation risk (MLCV-Risk) method** in this thesis, see Chapter 4. Tanner and Wong (1984) introduced this method for continuous time kernel hazard bandwidth estimation. It provides a good alternative to the above least squares methods when the computation cost is too high and produce better bandwidth estimates than the naïve bandwidths.

3.4 Grouped Time Kernel Hazard Estimation Framework

In economics it is more common to observe the duration of the spells in intervals of either equal lengths or unequal lengths than to observe the duration in continuous time. The transition in the process runs in the continuous time, but the data are taken at points in the time that known to the researcher, i.e in intervals. The survival function or the distribution function of the continuous time are used to construct the functions of the grouped time model. The basic assumption is that the survival probability does not change as a result of the way that the data is taken. Which means that the survival function at the time index points is the same as the survival probabilities of the continuous time. The time index points could be defined at the beginnings or at the ends of the intervals. This does not, theoretically, affect the construction of the grouped time model. We follow the design of Jenkins (2005) in this research by using the ends of the intervals as time index points, since this matches the grouping scheme in the R software that used in this research. The number of transitions inside each interval is the variable that used to construct the discrete time hazard estimators. In contrast to the continuous time models, in discrete time models the sequence of the transitions inside the intervals does not matter, although that the sequence of the intervals must be carefully considered. Accordingly, the problem of tie observations diminishes automatically with the grouping scheme.

The estimation of the grouped time hazard rate, also known as the **discrete time hazard** and **interval hazard**, by kernel estimation method is rarely considered in the nonparametric estimation literature. Tutz and Pritscher (1996) suggest a kernel estimator for the Life-Tables, and argue that to smooth the duration time an ordered variable kernel function can be used. We will proof that choosing this type of kernel functions is the right choice to smooth the duration time in the grouped time hazard estimator, and we will use this argument to develop the discrete time external kernel hazard estimator.

3.4.1 The Grouping Scheme

The grouping scheme provides sufficient information about the relationship between the hazard rate and the survival function in the continuous time and the discrete time. However, some functions in the grouped time, like the discrete time cumulative hazard function, may remain not fully specified without restrictive assumptions. This may cause an unidentification problem in the parametric estimation of the discrete time PH hazard model, where the coefficients of the discrete time variable may become interpretable under very narrow assumptions only. Jenkins (2005, Chapter 3, Section 3.3) shows an example of this case for a grouped Weibull variable. More information related to this issue is provided in Appendix D. To illustrate the grouping scheme and the relationships between the hazard and the survival functions we will use an equal length intervals grouping system, since this makes the discussion easier. We will show also that the construction under equal interval length is crucial for the development of the kernel hazard estimator.

Suppose that the range $(0, \infty)$ of the continuous time Y^c, with a continuous density function $f_{Y^c}(Y^c)$, is divided into B equal intervals having an equal length b. The intervals are numbered subsequently from 1 to B, and the time index is at the end of the intervals. So the intervals are defined as follows, $(0, b]$, $(b, 2b]$, ..., and $((B-1)\,b, Bb]$, (Bb, ∞), which are also written as $(0, y_1]$, $(y_1, y_2]$, $\ldots$, $(y_{B-1}, y_B]$ and (y_B, ∞). After the last observed complete interval, $(y_{B-1}, y_B]$, in the sample there is no upper bound, where $y^c_{B+1} = \infty$, and the hazard at the interval $(y^c_B, \infty]$ equals 1 irrespective of the interval length or the number of intervals. In practice the sample end date is y^c_B and the observations that remain in the state until y^c_B are right censored. We assume that the probability of survival after y^c_B is very low, $Pr(Y^c > y^c_B) \approx 0$, in our discussion. So, the reminder probability of f_{Y^c} could be ignored when the variable is grouped. This is consistent with the behaviours of the transition processes in economics if the sample period was sufficiently long.

Let the positive integer J^o denote the number of the interval at with the transition occurs. J^o is denoted also as the **episode** number, where $j^o = 1, 2, 3, \cdots, B$. Then $j^o = ((j^o - 1)\,b, j^o b]$ is the interval j^o written using in the interval length form to denote the time index points, and $j^o = \left(y^c_{j^o-1}, y^c_{j^o}\right]$ is the same interval with the time index points written in Y^c form. Then $y^c_1, y^c_2, \cdots, y^c_B$ are the time index points that associated with the intervals $1, 2, 3, ..., B$ respectively[3]. Note that $y^c_{B+1} = \infty$ and the hazard in the episode $(y^c_B, \infty]$ equals 1 irrespective of the interval length or the number of intervals.

Let $f(j^o) = f_{j^o}$ be the probability mass of the discreteised time on episode j^o, then the distribution function for Y^c and the distribution function of j^o must be equal at the

[3]Clearly: $y^c_1 = b$, $y^c_2 = 2b$, $\cdots$, $y^c_B = Bb$.

ends of the intervals by construction, then the two functions are related as follows:

$$\sum_{r=1}^{j^o} f_{J^o}(r) = \int_0^{y^c_{j^o}} f_{Y^c}(u)du = F_{Y^c}(y_{j^o}), \tag{3.23}$$

where the summation is for the discretized distribution function and the integration is for the continuous distribution function[4]. This shows how the matching between the area the under distribution functions of the continuous variable and the discretized variable is specified mathematically. The survival probability in a grouped time model in interval j^o is defined as the probability of survival the interval and beyond, which is the complement of the distribution function at the end of the previous interval, so that $S_j = 1 - F_{Y^c}\left(y^c_{j^o-1}\right)$.

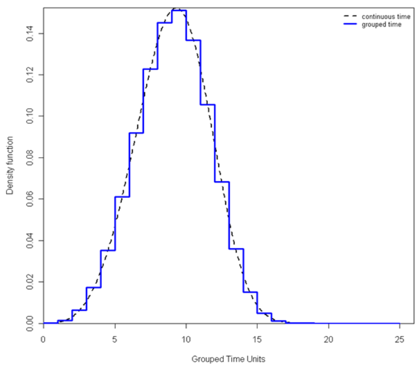
Figure 3.1 illustrates the grouping scheme using an example of a Weibull continuous transition variable with shape parameter 4 and scale 10. The Weibull density at this shape parameter approximates the normal density, see Mood and Graybill (1963) and Spanos (1999), and the scale parameter is chosen arbitrarily to suit the interval length which set to be $b = 1$. The plots in the figure show that the density, the distribution and the survival functions of the continuous Weibull and the grouped Weibull are very closely matched, but the hazard rate, which has an upward slope because the shape is larger than 1, does not match. Mathematically in this example of Weibull distribution, the hazards match only when the scale tend to ∞ or when $b \longrightarrow 0$.

When the model include the right censoring problem, the censoring runs as a continuous time process (on the underlying continuous time). The observation is right censored if the right censoring continuous time, C^c, is less than the continuous transition time Y^c, so that we cannot observe the transition time. Generally, observation i is right censored in episode j^o if $C^c_i < Y^c_i$ and $C^c_i \in ((j^o - 1)\, b, j^o b]$. A right censoring indicator is defined in the reverse relationship

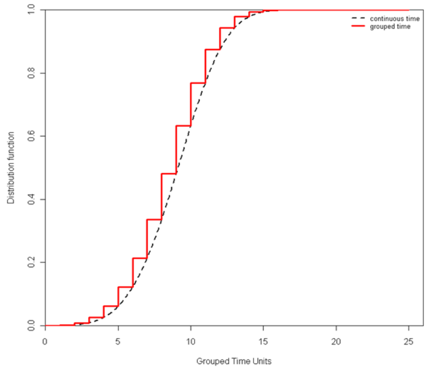
$$\triangle_{ij^o} = I\left(Y^c_i \leq C^c_i, (j^o - 1)\, b < Y^c_i \leq j^o b\right). \tag{3.24}$$

The sample of the grouped time hazard contains $\{J^o_i, \triangle_i, \mathbf{X}_i\}^n_{i=1}$, where J^o_i is a scalar that indicates the number of episodes that individual i spent in the state. So, J^o_i is a positive integer number for all i. The underlying continuous time in the process is no longer Y^c, because of the existence of the right censoring problem, instead it is $T^c_i = min(Y^c_i, C^c_i)$. So the underlying continuous time process is identical to the process that described in the previous section, but the hazard can only be estimated for the intervals. The censoring indicator function, $\triangle_i$, takes a value 1 if observation i is uncensored and the vector $\mathbf{X}_i$ is a vector of mixed independent variables.

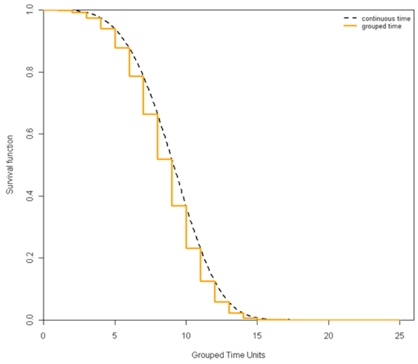
[4]For simplicity we are not using a notation for the cumulative distribution function of the discretized variable.



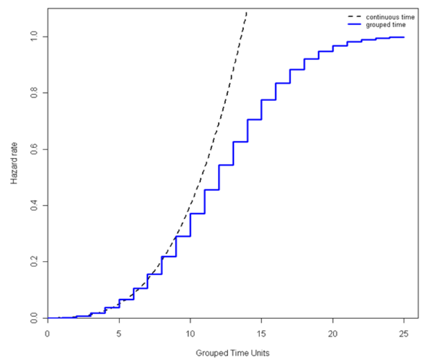
The density function



The distribution function



The survival function



The hazard function

Figure 3.1: Grouping Scheme of Weibull with shape 4 and scale 10, using interval length 1.

3.4.2 Smoothing Methods for the Grouped Time Variable

How to smooth a discretised continuous random variable was considered early in the development of the kernel smoothing techniques, for example in Wand and Jones (1995) and Simonoff (1996). We will describe two methods, the first smooths the underlying continuous variable with the continuous variables smoothing technique. The second method tries to smooth the discrete grouped variable.

Let the vector $\mathbf{p}^T = [f_1, f_2, .., f_B]$ be a vector of length B contains the probabilities of the intervals, where for interval j^o the probability is $f_{j^o} = \int\limits_{(j^o-1)/B}^{j^o/B} f_{Y^c}(u)\,du$. Let $\widehat{f}_{j^o}$ be the estimated probability for interval j^o and d_{j^o} be the number observations in interval j^o. The estimator $\widehat{f}_{j^o}$ is consistent if $\widehat{f}_{j^o} \longrightarrow f_{j^o}$ as $d_{j^o} \longrightarrow \infty$ and $n \longrightarrow \infty$, see Simonoff (1996, Chapter 6). Then estimating the density for the discretised continuous variable, j^o, by the relative frequency $\frac{d_{j^o}}{n}$ is inappropriate because this fails to use the information about the density of the continuous variable f_{Y^c} outside the interval j^o. In other words, the estimated probability of the discrete function $\widehat{f}_{j^o}$ must use the information from all the intervals, which means that it should be a function of all the probabilities in $\mathbf{p}$. The probability f_{j^o} can then be specified as a function on the probabilities of all other intervals, f_{-j^o}, particularly on the nearest neighbours to j^o. So that the estimated density $\widehat{f}_{j^o}$ should be estimated so that for any nearby interval $j^{o\prime}$, with $j^o \neq j^{o\prime}$ the estimates must satisfy $\widehat{f}_{j^{o\prime}} \longrightarrow \widehat{f}_{j^o}$ as $j^{o\prime} \longrightarrow j^o$.

Smoothing the grouped time using the nonparametric regression technique

The smoothing of the grouped variable, j^o, is possible by the nonparametric regression technique as follows; Simonoff (1996, page 219):

$$\widehat{f}_{j^o} = \frac{\sum_{j^o=1}^{B} k\left(\frac{j^o/B-(j^o-1)/B}{h}\right)\frac{d_{j^o}}{n}}{\sum_{j^o=1}^{B} k\left(\frac{j^o/B-(j^o-1)/B}{h}\right)}. \tag{3.25}$$

The disadvantage of this method is that it suffers from the Boundary Bias. Where the continuous variable kernel function that used in Eq (3.25) smooths negative values in the first episode, as in the external kernel hazard estimator in the continuous time model. Some authors recommend using of local linear regression or local polynomial nonparametric regression to reduce the bias in the discretised density estimator. However, this can only partially remove it. This method is not suitable for smoothing the grouped time variable in the discrete kernel hazard estimator. Since it has the above disadvantage and it is complicated to be transformed to construct a grouped time hazard estimator that has a similar form of the continuous time kernel hazard estimators.

Smoothing the grouped time with the Wang and Van Ryzin discrete ordered kernel function

The second approach uses a discrete ordered kernel function directly to smooth the grouped time variable. The development of the ordered variable kernel function is illustrated in Titterington (1980). The method uses assumptions about the probabilities in $\mathbf{p}$ and an estimator of the form:

$$\widehat{\mathbf{p}}\left(\gamma\right) = \left(1-\gamma\right)\mathbf{d} + \gamma\mathbf{p}, \quad \gamma \in \left[0,1\right], \tag{3.26}$$

where γ is estimated by a data- driven method. The information about the probabilities in all intervals is used to estimate the probability in interval j^o. Different distribution assumption are used for the probabilities in $\mathbf{p}$. One can think of $\mathbf{p}$ as probabilities from: a Multinomial, Uniform, Poisson,... discrete distribution. For duration models with single spell one can think of the transition as a single event that takes place at any of the episodes $1, 2, ..., B$. Then, if the transition is observed at interval j^o, it can be considered as a success after $j^o - 1$ failures. Therefore, one can think about the probabilities in $\mathbf{p}$ as Geometric distribution generated probabilities [5]. Wang and Van Ryzin (1981) use

[5] Titterington (1980) shows that when a multinomial distribution is assumed for the probabilities in $\mathbf{p}$, the formula produces the Aitchison and Aitken (1976) kernel function for unordered variables.

an estimator of this form to develop the so-called the Geometric Kernel function for discrete ordered variables. Accordingly, the assumption that the intervals are all equal length is crucial to smooth the grouped time variable because otherwise, the assumption about the probabilities in $\mathbf{p}$ has to be changed, in other words, another discrete kernel function is required.

The Geometric Kernel function is symmetric and takes the following form:

$$l\left(J_i, j, \gamma\right) = \begin{cases} 1-\gamma & \text{if } \left|J_i^o - j^o\right| = 0 \\ \frac{1}{2}\left(1-\gamma\right)\gamma^{\left|J_i^o - j^o\right|} & \text{if } \left|J_i^o - j^o\right| \neq 0 \end{cases}, \tag{3.27}$$

where $l\left(\cdot,\cdot,\cdot\right)$ depends on the smoothing parameter (the bandwidth) γ and the number of the intervals B. The ordered kernel gives B local weights for the difference between the point of smoothing j^o and the observations in the sample, J_i^o. The weight $(1-\gamma)$ is the highest and is given if the difference is zero, i.e. the observations in the sample that are in the same interval with j^o are given higher weights by the kernel function than the observations in other intervals. The wider the difference between j^o and J_i^o the lower the weight that is given to J_i^o by the kernel function. This discrete ordered kernel function exploits the information from all intervals in the discretised density to estimate the probabilities, the estimated probabilities have the quality that for any two intervals j^o and $j^{o\prime}$, with $j^o \neq j^{o\prime}$, $\widehat{f}_{j^{o\prime}} \longrightarrow \widehat{f}_{j^o}$ as $j^{o\prime} \longrightarrow j^o$, which is the required property in the discretised density estimator.

The properties and the consistency conditions for the discrete ordered kernel function are shown by Wang and Van Ryzin (1981), Ahmad and Cerrito (1994) as follows:

1 The weights of any interval in $\{1, 2, 3, \ldots, B\}$ must be proper probabilities and add up to 1;

$$\sum_{j^o=1}^{B} l\left(., j^o, \gamma\right) = 1, \tag{3.28}$$

and $l\left(., j^o, \gamma\right) \geq 0$, for any j^o.

2 $\gamma \rightarrow 0$ as $n \rightarrow \infty$.

3 $\gamma/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$, and $l\left(J_i, j, \gamma\right)$ has a continuous first derivative.

So, the Geometric kernel has similar asymptotic properties to the continuous variables kernel functions.

The properties of the ordered variable kernel function are given by Tutz and Pritscher (1996). The discrete ordered kernel function has an advantage that it does not smooth negative values in the first interval. So, the source of the Boundary Bias is eliminated from the grouped time kernel hazard estimator.

3.4.3 Discrete Time Unconditional Kernel Hazard

The grouped time hazard rate is constructed from the continuous time survival function by making the grouped time and the continuous time survival functions equal at the ends of the intervals. Let $\theta\left(j^{o}\right)$ be the hazard rate of the grouped time on interval j^{o}, Jenkins (2005) and Brecht and Brecht (1996) show that the survival probability of the discrete time hazard is:

$$\begin{aligned} S_{Y^c}\left(y^c_{j^o}\right) &= \Pr\left(Y^c > y^c_{j^o-1}\right), & (3.29)\\ &= \prod_{r=1}^{j^o}\left(1-\theta\left(r\right)\right). & (3.30)\end{aligned}$$

Where the grouped time hazard is defined as

$$\begin{aligned} \theta\left(j^{o}\right) &= \Pr\left(y^c_{j^o-1} < Y^c \leq y^c_{j^o}\left|Y^c > y_{j^o-1}\right.\right), & (3.31)\\ &= \frac{\Pr\left(y^c_{j^o-1} < Y^c \leq y^c_{j^o}\right)}{\Pr\left(Y^c > y^c_{j^o-1}\right)}, & (3.32)\\ &= 1-\frac{S_{Y^c}\left(y^c_{j^o}\right)}{S_{Y^c}\left(y^c_{j^o-1}\right)}, & (3.33)\end{aligned}$$

and the density of the grouped time variable can be written as a function of the hazard and the survival functions as follows:

$$f\left(j^{o}\right)=\frac{\theta\left(j^{o}\right)}{1-\theta\left(j^{o}\right)}\prod_{r=1}^{j^o}\left[1-\theta\left(r\right)\right]. \qquad (3.34)$$

Brecht and Brecht (1996) suggests a product limit estimator of the discrete time survival function which takes the following form [6]:

$$\begin{aligned} \widehat{S}_n\left(j^{o}\right) &= \prod_{r=1}^{j^o}\left(1-\frac{d_r}{n_r}\right).I\left(n_{j^o}>0\right), \\ &= \prod_{r=1}^{j^o}\left(1-\widehat{\theta}_n\left(r\right)\right).I\left(n_{j^o}>0\right), & (3.35)\end{aligned}$$

where $\widehat{\theta}_n\left(j^{o}\right)=\frac{d_{j^o}}{n_{j^o}}$,

$$\begin{aligned} d_{j^o} &= \sum_{i=1}^{n} I\left(J^o_i \leq j^o; \triangle_i = 1\right) - \sum_{i=1}^{n} I\left(J^o_i \leq j^o - 1; \triangle_i = 1\right), & (3.36)\\ &= \sum_{i=1}^{n} I\left(J^o_i = j^o; \triangle_i = 1\right), & (3.37)\end{aligned}$$

[6] We moved the time index on Brecht and Brecht (1996) estimators to the end of the intervals to make the notation consistent with our equations.

and $n_{j^o} = \sum_{i=1}^{n} I\left(J_i^o \geq j^o\right) = n - \sum_{i=1}^{n} I\left(J_i^o \leq j^o - 1\right)$ is the number at risk at the end of episode j^o (just after the time index point y_{j-1}^c). The estimated cumulative hazard from the data is

$$\widehat{\Lambda}_n\left(j^o\right) = \sum_{r=1}^{j^o} \widehat{\theta}_n\left(r\right).I\left(n_r > 0\right), \tag{3.38}$$

To develop the external kernel hazard estimator in the grouped time we start from the formula of Watson and Leadbetter (1964a) in Eq (3.3). Replacing the continuous variable kernel function $k\left(\cdot\right)$ by the discrete ordered variable kernel function $l\left(.,.,\gamma\right)$ and the integration by a summation over all the intervals, the formula becomes

$$\begin{aligned} \theta\left(j^o\right) &= \sum_{r \in B} l\left(r, j^o, \gamma\right) \frac{\triangle F_n\left(r-1\right)}{1 - F_n\left(r-1\right)}, & (3.39) \\ &= \sum_{r \in B} l\left(r, j^o, \gamma\right) \frac{F_n\left(r\right) - F_n\left(r-1\right)}{1 - F_n\left(r-1\right)}, & (3.40) \end{aligned}$$

where $F_n\left(\cdot\right)$ is the sample distribution function of the grouped time.

Inside the interval the process runs in the continuous time,

$$\begin{aligned} F\left(j^o\right) - F\left(j^o - 1\right) &= \int_{y_{j^o-1}^c}^{y_{j^o}^c} f_{Y^c}(u) du, & (3.41) \\ &= f\left(j^o\right). & (3.42) \end{aligned}$$

Then for interval j^o the density is estimated using the sample distribution function as follows:

$$\begin{aligned} \widehat{F}_n\left(j^o\right) - \widehat{F}_n\left(j^o - 1\right) &= n^{-1} \sum_{i=1}^{n} I\left(J_{(i)}^o = j^o\right), & (3.43) \\ &= \frac{d_{j^o}}{n} = \widehat{f}_n\left(J^o\right). & (3.44) \end{aligned}$$

Then

$$\widehat{\theta}\left(j^o\right) = \sum_{r \in B} l\left(r, j^o, \widehat{\gamma}\right) \frac{\widehat{f}_n\left(j^o\right)}{1 - \widehat{F}_n\left(r-1\right)}. \tag{3.45}$$

From the sample this produces the following external grouped time hazard estimator, after rearranging the terms:

$$\begin{aligned} \widehat{\theta}\left(j^o\right) &= \sum_{r=1}^{B} \frac{l_0\left(r, j^o, \widehat{\gamma}\right) \sum_{i=1}^{n} I\left(J_i^o = r\right)}{n - \sum_{i=1}^{n} I\left(J_i^o \leq r - 1\right)}, \\ &= \sum_{i=1}^{n} \frac{l_0\left(J_i^o, j^o, \widehat{\gamma}\right)}{n_{J_i^o}}. & (3.46) \end{aligned}$$

This is very similar to the external hazard estimator for continuous time models in Eq (3.7). In this equation $n_{J_i^c}$ is the size of the risk set at the beginning of episode J_i^o, and is equivalent to $n-i+1$ in the continuous estimator, which is the size of the risk set just after the distinct failure time $T_{(i)}$.

For right censored data, the external hazard estimator includes a censoring indicator as follows,

$$\widehat{\theta}^{EKH}(j^o) = \sum_{i=1}^{n} \frac{l(J_i^o, j^o, \widehat{\gamma}) \triangle_i}{n_{J_i^o}}. \tag{3.47}$$

We call this estimator the **discrete time unconditional external kernel hazard (EKH) estimator**.

3.4.4 Discrete Time Conditional Kernel Hazard Estimator

To estimate the conditional hazard rate one can use an estimator that directly extended from the continuous time external kernel hazard estimator. The conditional kernel hazard estimator of Beran (1981) in Eq (3.12) does not need modification because we can take the difference in the numerator using the grouped time index points, i.e. at the ends of the episodes. But to extend the external kernel hazard in Eq (3.19), one can replace the continuous variable kernel by the Geometric kernel function. The grouped time estimators are intended to be used with a mixed independent variables, which affects the smoothing of the sub-survival (sub-distribution) function in the formula. The distribution function at the ends of the intervals of the grouped time, j^o, is estimated by the conditional mean regression function estimator of the CDF of the continuous time variable Y^c at y_{j^o}. In Li and Racine (2007, Chapter 6, page 182), the estimator takes the form

$$\begin{aligned}
\widehat{F}_{Y^c;\mathbf{h}}\left(y_{j^o}^c \,|\mathbf{x}\right) &= \sum_{i=1}^{n} \frac{W(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}}) I\left(T_i^c \le y_{j^o}\right)}{\sum_{i=1}^{n} W(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}})}, &(3.48)\\
&= \sum_{i=1}^{n} \mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}}) I\left(T_i^c \le y_{j^o}^c\right), &(3.49)\\
&= \sum_{i=1}^{n} \mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}}) I\left(J_i^o \le j^o\right), &(3.50)\\
&= \sum_{i: J_i \le j^o} \mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}}), &(3.51)\\
&= \widehat{F}_{\mathbf{h},2}\left(j^o \,|\mathbf{x}\right). &(3.52)
\end{aligned}$$

Although T_i^c in Equations (3.48) and (3.49) is not observed but by construction

$$\widehat{F}_{Y^c;\mathbf{h}}\left(y_{j^o}^c \,|\mathbf{x}\right) = \widehat{F}_{\mathbf{h}}\left(j^o \,|\mathbf{x}\right), \tag{3.53}$$

the construction makes $I\left(T_i^c \le y_{j^o}^c\right) = I\left(J_i^o \le j^o\right)$, which is used in Eq (3.50). The regression function can then be written in a sub-distribution function form. The estimator estimates the distribution function conditional on a mixed independent variables, so it can be used directly to estimate the sub-distribution functions at the ends of the intervals.

The survival function is estimated by the same equation but with the inverse strict inequality sign as follows:

$$\begin{aligned}
\widehat{S}_{Y^c;\mathbf{h}}\left(y_{j^o-1}\left|\mathbf{x}\right.\right) &= \sum_{i=1}^{n} \frac{W(\mathbf{x},\mathbf{X}_i,\widehat{\mathbf{h}}) I\left(T_i^c > y_{j^o-1}\right)}{\sum\limits_{i=1}^{n} W(\mathbf{x},\mathbf{X}_i,\widehat{\mathbf{h}})}, &(3.54)\\
&= \sum_{i:J_i > j^o-1} \mathbf{A}(\mathbf{x},\mathbf{X}_i,\widehat{\mathbf{h}}), &(3.55)\\
&= \sum_{i:J_i \ge j^o} \mathbf{A}(\mathbf{x},\mathbf{X}_i,\widehat{\mathbf{h}}), &(3.56)\\
&= \widehat{S}_{\mathbf{h},2}\left(j^o\left|\mathbf{x}\right.\right), &(3.57)
\end{aligned}$$

which estimates the conditional survival function at the beginning of interval j^o. This sub-survival function is the complement of the sub-distribution function at the previous episode, ,$1-\widehat{F}_{\mathbf{h},2}\left(j^o-1\left|\mathbf{x}\right.\right)$.

Then Beran (1981) conditional kernel hazard estimator in the grouped time and mixed independent variables is

$$\begin{aligned}
\widehat{\theta}^{BKH}\left(j^o\left|\mathbf{x}\right.\right) &= \sum_{i=1}^{n} \frac{\widehat{S}_{\widehat{\mathbf{h}},1}\left(j^o\left|\mathbf{x}\right.\right) - \widehat{S}_{\widehat{\mathbf{h}},1}\left(j^o+1\left|\mathbf{x}\right.\right)}{\widehat{S}_{\widehat{\mathbf{h}},2}\left(j^o\right)}, &(3.58)\\
&= \frac{\sum\limits_{i:J_i^o=j^o} \mathbf{A}(\mathbf{x},\mathbf{X}_i,\widehat{\mathbf{h}})\ \triangle_i}{\sum\limits_{i:J_i^o\ge j^o} \mathbf{A}(\mathbf{x},\mathbf{X}_i,\widehat{\mathbf{h}})}, &(3.59)
\end{aligned}$$

where

$$\widehat{S}_{\widehat{\mathbf{h}},1}\left(j^o\left|\mathbf{x}\right.\right) = \sum_{i=1}^{n} \frac{W(\mathbf{x},\mathbf{X}_i,\widehat{\mathbf{h}}) I\left(J_i^o \ge j^o; \triangle_i = 1\right)}{\sum\limits_{i=1}^{n} W(\mathbf{x},\mathbf{X}_i,\widehat{\mathbf{h}})},$$

The conditional external kernel hazard estimator for continuous time from Eq (3.19) is extended to the grouped time as follows

$$\widehat{\theta}^{CEKH}\left(j^o\left|\mathbf{x}\right.\right) = \sum_{i=1}^{n} l\left(J_i^o, j^o, \widehat{\gamma}_0\right) \frac{\mathbf{A}(\mathbf{x},\mathbf{X}_i,\widehat{\mathbf{h}})\ \triangle_i}{1-\sum\limits_{i:J_i<j} \mathbf{A}(\mathbf{x},\mathbf{X}_i,\widehat{\mathbf{h}})}, \tag{3.60}$$

This estimator is denoted as the **grouped time conditional external kernel hazard (CEKH) estimator**. As with the continuous time conditional external conditional hazard estimator, in the grouped time hazard if the Nadaraya-Watson weights in Eq (3.60) were replaced by n^{-1} the formula reduces to the discrete time unconditional external hazard estimator in Eq (3.47).

3.4.5 The Estimation of the Bandwidths of the Grouped Time Kernel Hazard Estimators

Since the grouped time kernel hazard estimators are straightforward extensions of the continuous time kernel hazard estimators, one claims that the methods to estimate the bandwidths might not be different. However, to suggest a method that is based on the bias-variance tradeoff, we need to find the bias and the variance formulas of the grouped time estimator. Which are the equivalent formulas to the bias and variance formulas of the continuous time kernel hazard estimator that shown in the equations from Eq (3.8) to Eq (3.10). However, this work is beyond the level of this research and it is not attempted.

A method of bandwidths estimation that tradeoff between bias and variance in the discrete time kernel hazard estimator might produce an estimation method work similarly to the least squares cross-validation method of Rudemo (1982) in discrete variable kernel density estimation. The bias-variance tradeoff method generally aims to find the optimum bandwidths, the bandwidths that are not large so they over-smooth the kernel estimator (increase the variance of the kernel estimator), and not small so they under-smooth the kernel estimator (increase the bias). The optimum bandwidths in this sense could be achieved by optimizing a cross-validation risk function of the form $R\left\{\theta\left(j^o\,|\mathbf{x}\right), \widehat{\theta}^{(\cdot)}\left(j^o\,|\mathbf{x}\right)\right\}$, see Wasserman (2006, Chapter 6). An objective function of this form could be a function that minimize the integrated squared risk function between the kernel hazard and the true hazard, which could produce an objective function similar to the objective function of the least-square cross validation method, this method in known as the least-square cross validation risk method[7].

We suggest Maximum Likelihood Cross-Validation Risk (MLCV-Risk) method, with an objective function that is based on **kullback-leibler loss function**. It is applied in the continuous time kernel hazard estimation approach in Tanner and Wong (1984). The leave-one-out method is used in the discrete time hazard, so that the maximum likelihood function of the grouped time hazard that described by Jenkins (2005) can be rewritten as MLCV-Risk objective function as follows:

$$ML\left(\gamma_0, \mathbf{h}\right) = \sum_{i=1}^{n}\left[\triangle_i \log\left(\frac{\widehat{\theta}_{-i}^{(\cdot)}\left(J_i^o\,|\mathbf{x}_i\right)}{1-\widehat{\theta}_{-i}^{(\cdot)}\left(J_i^o\,|\mathbf{x}_i\right)}\right) + \sum_{r=1}^{J_i^o}\log\left(1-\widehat{\theta}_{-i}^{(\cdot)}\left(r\,|\mathbf{x}_i\right)\right)\right] \tag{3.61}$$

[7]We use the term cross validation risk function to refer to the cross validation objective function when the estimated kernel function is a hazard function, Wasserman (2006) shows that both names can be used to refer to the CV-objective function. We think this may help the reader to distinguish the case when the CV method is applied to the density (minimizes a mean square error) from the case when the CV is applied to the hazard (minimize a squared risk function).

where $\widehat{\theta}_{-i}^{(\cdot)}$ is the kernel hazard estimated with the i^{th} observation left out, $(\cdot)$ denotes either BKH or CEKH.

The MLCV method has a traditional problem with the fat tail and monotonicity increasing or decreasing functions, where it over-smooths the bandwidths and increases the variance in the estimator. So, the kernel hazard estimator may get over-smoothed easily. Accordingly, more research is needed in this topic.

A less attractive choice is to use optimum cross validation bandwidths of kernel conditional probability mass function estimate as naïve bandwidths in the hazard estimator. The only advantage of the naïve bandwidths is that they are more attractive computationally than CV risk bandwidths. Unfortunately, we are forced to use naïve bandwidths in the simulation study due to the tough commutation cost of MLCV-Risk method.

3.5 Monte Carlo simulations

A Monet Carlo simulation study is designed to examine the discrete time kernel hazard estimators. The study compares between the kernel estimators and parametric PH estimators of the grouped time hazard rate. The development of the **data generating process (DGP)** uses the PH framework: we simulate a continuous time duration variable from Weibull distribution function and then apply the grouping scheme. So, the baseline hazard is specified in the simulation in the continuous time and the discrete grouped time. The observed heterogeneity function is defined as a function that depends on time invariant independent variables only. The continuous duration time is grouped using an equal interval length that fixed at $b = 1$ in all simulations. The functional form of the baseline hazard depends on the parameters of the Weibull distribution. The true baseline hazard and the coefficients of the independent variables, i.e. the contribution of the independent variables to the baseline hazard, are designed to be known in both continuous and grouped time, so the distance between the predicted hazard and the true hazard is measurable in the DGP.

A fully parametric accelerating failure time AFT model is estimated for the continuous time hazard model, before the grouping of the time. The relationship between the coefficients of the independent variables in the AFT form with the coefficients of the PH with Weibull baseline hazards is illustrated in Jenkins (2005). The coefficients of the independent variables in the continuous time PH form and the discrete time PH form are the same, and the functional forms of the baseline hazard in the continuous time and the discrete time is mapped though the survival function of the continuous time. Because the coefficients are the same in the continuous time PH and the discrete time PH, only the discrete time model is estimated. Additionally, because the kernel estimators in the continuous time have Boundary Bias problem and they need hard computation, they are not included in the simulation study. The objective of including the continuous time AFT model in the simulation, which is the only continuous time estimator that included, is to examine whether we simulated the proportional effect of the independent variables correct or not.

Jenkins (2005) provides sufficient and easy to follow guidance to design the simulations, where the functional form of the Weibull AFT regression model, Weibull continuous time PH model, grouped time Weibull PH model and the parametrizeation of the coefficients of the independent variables between AFT and PH models is clearly illustrated.

The DGP of Baker and Melino (2000) and Nicoletti and Rondinelli (2010), which are the only simulation studies that we are able to find for discrete time conditional hazard in economic, are inappropriate to use with the kernel estimators. Baker and Melino (2000) and Nicoletti and Rondinelli (2010) simulate a discrete hazard directly with no means to follow the process in the continuous time. The distance between the true hazard and the predicted hazards is not measurable directly, where the DGP is parametrically designed, which is the reason that makes the kernel estimators cannot be included easily. The assessment of the performance in Baker and Melino (2000) and Nicoletti and Rondinelli (2010) simulation studies is driven for the parametric model on the bases of the estimated coefficient values only. A detailed description of the functional form of the baseline hazard and the observed heterogeneity function in the DGP is provided, however, we still find that the modification of their DGP to accommodate the kernel estimator needs as much effort as that used to develop our original DGP that described below.

3.5.1 The Design of the Data Generating process

The development of the simulation study in this research uses a similar technique to that developed in the simulation studies in Hess et al. (1999), Cao and Jácome (2004), Dikta et al. (2005), Jácome and Iglesias-Pérez (2008), Cho and White (2010) and Cho et al. (2011). However, we introduce the time grouping scheme to the DGP to allow for estimating the grouped time hazard. The competing risks random right censoring is simulated along with the simulated data. The distribution of the right censoring time is continuous and defined as a $Uniform(0, u_\pi)$ random variable. So the right censoring is not duration dependent and the censoring time is simulated independently from the transition time and the independent variables. The assumption that the censoring time is Uniform is easier to handle than assuming that it follows another Weibull distribution, because one can then control the proportion of the right-censored observations, π, in the Monte Carlo samples better.

The right censored observations are the observations that correspond to the negative value in the variable $C^c - Y^c$, where Y^c is the transition time and C^c is the right censoring time. Then we can simulate from the distribution function of $C^c - Y^c$ and set the proportion of the right censored observations, which is $\pi = F\left(c^c - y^c < 0\right)$. The upper limit of the censoring time, u_π, is approximated for each given parameters of Weibull distribution. The numerical approximation method approximates the population distribution function $F(c^c - y^c < 0)$, and is available in the package `distr` in R software. The value of π is determined arbitrarily in the study.

Two DGPs are studied, the first for the study of the unconditional hazard estimators, without the independent variables simulated in the study. The second DGP simulation study is for the conditional hazard estimators with time invariant independent variables only. There are two independent variables that are included, a continuous variable x^c and a categorical variable x^d with only two possible outcomes $[0, 1]$. The R software is used to do all the simulations, with the codes to estimate the MLCV naive bandwidths written in the simulation program. The execution time of the simulation program varies with the values of Weibull parameters and the proportion of the right censored observations, the total of the execution times for all the simulations exceeded the 1500 hours.

Let λ be the scale parameter and k be the shape parameter of the continuous Weibull transition time, and that π is used to determine the value of u_π. In each sample, q, in the unconditional hazard DGP, the variables of the i^{th} observation are simulated as follows:

1. Simulate Y_i^c from f_{Y^c}, a $Weibull(scale = \lambda, shape = k)$ continuous density.

2. Simulate C_i^c from r_{C^c}, the $Uniform(0, u_\pi)$ continuous density.

3. Compute $T_i^c = min(Y_i^c, C_i^c)$, and $\triangle_i = I(Y_i^c \leq C_i^c)$.

4. Bin T_i^c into equal groups each of length $b = 1$.

5. Apply the estimators of the discrete time hazard for the grouped transition model and the continuous time hazard AFT estimator to the continuous time variable.

6. Compute the measures of distance for the predicted hazard rate from the true grouped time hazard, the measure of distance is applied for the grouped time hazard estimators only.

The parameter values that chosen for the Weibull distribution are $\lambda \in \lambda = [1, 3]$, and $k \in \mathbf{k} = [1, 0.25, 0.50, 0.75]$. The proportions of the censored observation are chosen from the set $\pi \in [0\%, 2\%, 5\%, 10\%, 25\%]$. The distributions are simulated for the sample sizes 100, 200, 500 and 1000 individuals. The number of the Monte Carlo samples repetitions is 500 in all simulations.

The discrete time PH models are estimated using Jenkins (1995) method, which include the following steps:

1. Pool the sample into an unbalance panel data shape, for each individual, i, the time variable is a discrete number that takes values from one to the number of the last episode that observed to that individual, J_i^o. The time variable in the pooled sample is known as **the elapsed time** in the state a_{ir}, where $i = 1, 2, \cdots, n$ and $r = 1, 2, \cdots, j_i^o$. Then for each individual in the sample we generate a number

of rows that equivalent to the number of the episodes that are observed to that individual in the state.

2. The independent variables for each individual are repeated in each row until the last row, which is the row that is associated with the last episode that observed for that individual. Where the independent variables, in our simulation, are all time invariant, so they do not change with the elapsed time in the state.

3. Introduce a dummy variable, d_{ir}^{d}, in the pooled sample that takes one on the last episodes only and only if the individual in the sample made a transition. So, the sum of d_{ir}^{d} in the pooled sample gives the total number of transitions. We will refer to this variable as the **pooled sample transition dummy**.

4. Use the binary choice modelling technique to estimate a binary model for the dummy variable d_{ir}^{d} using a **complementary log-log (cloglog) link function**. The binary choice regression is regressed on the elapsed time variable a_{ir} and the independent variables.

5. The predicted probabilities of the complementary log-log binary model are directly the estimates of the discrete time hazard rate. The predictions at the last episode, j_i^o, give the predicted hazard at the last episode that observed for individual i.

A transformation of the elapsed time in the state variable, a_{ir}, is needed in the pooled sample binary model to define the functional form of the baseline hazard in the parametric PH estimator. Using $log\left(a_{ir}\right)$ transformation estimates the Weibull baseline hazard directly in the cloglog model, but we cannot identify the Weibull parameters if the Weibull shape parameters not equal 1, see the mathematical proofs in Appendix D. Using dummy variables instead of a_{ir} for each elapsed time unit, or grouped a_{ir}, is known as the semi-parametric specification of the baseline hazard, also as the **piecewise linear baseline hazard**. The grouped a_{ir} is used to handle the decreasing number of individuals in the state on the late episodes, see Jenkins (2005). The unconditional PH hazard estimator is obtained by regressing the transformed a_{ir} variable (or dummies) only.

For each Monte Carlo sample the above steps are used, where the sample sizes 100, 200, 500 and 1000 are pooled after simulating the variables. The size of the pooled sample depends also on the Weibull parameters and the proportion of the right censored observations. The estimation of the parametric PH model is harder if the pooled sample is too long. For the simulations with Weibull with scale parameter 3 and shape parameter 0.25, some spells were simulated very long, exceeded a 1000 episodes. Then the pooling of the sample and the estimation of the PH model became more tedious than the estimation of the nonparametric kernel hazard, where the size of some pooled sample exceed a million (individuals$\times$elapsed time unites). Accordingly, an upper limit of the binning process is used to reduce the time that needed to estimate the parametric PH model.

All the spells that exceeded 250 episodes are right censored in the simulation study, i.e we set $B = 250$ in all simulations. However, we attempted to make the value of B not effective to the predetermined right censoring proportion, π. The grouping scheme shows that all the observations in the interval $B + 1$ are right censored. Low value of B will increase the Monte Carlo sample proportion of the right-censored observations, $\widehat{\pi}$, to a larger value than π. Ignoring B in the simulation will lead to have samples with spells that are very long, particularly on the sample that simulated from Weibull density function with shape equals 0.25. Therefore, we choose $B = 250$, so that $Y_B = 250$ since the interval length is 1, which appeared to be a suitable choice. Where, it is easy to show that the survival probability for the continuous time at the end of episode B, $S\left(Y_{250}\right) = \exp\left[-\left(\frac{250}{\lambda}\right)^k\right]$, and it is very low, ignorable, when the shape parameter is 1, 0.75 and 0.50. For the sample that are simulated from Weibull with shape parameter 0.25, the survival probability at the end of episode B is 0.0188 and 0.0487 for the scale parameter 1 and 3 respectively, which increases the possibility that $\widehat{\pi}$ be larger than π.

Table 3.1 illustrates the mapping from the continuous Weibull transition variable to the grouped variable by showing some statistical measures and functions. The first panel shows the expected value, the median (50^{th}) quantile, the 95^{th} and the 99^{th} quantiles of the continuous time, the continuous time hazard rate at the index point of the first episode, and the continuous time survival function at the index point of the first episode and at each of the location measures. Panel B in the table shows the grouped time measures and functions. For each column of a statistical measure of the continuous time variable in Panel A we include a column in Panel B that shows the episode where that measure is contained in the grouped time variable when the interval length is $b = 1$. The Weibull with scale parameter $\lambda = 1$ and shape parameter $k = 1$, for example, has expected value equals $\lambda = 1$, which is the index point of the first interval in the grouped time. The simulations with scale 1 may be practically unsuitable, but theoretically they have nice properties that help to understand the process better.

As Table 3.1 shows, the hazard rate of the continuous time variable at the index point of the first episode equals the value of the shape parameter, $\phi_{Y^c}(1) = k$, and survival functions are all equal for all shape parameters, $S_{Y^c}(1) = 0.3679$. For the grouped time Weibull with scale parameter $\lambda = 1$, the discrete time hazard in the first interval is the same at all shape parameters. Accordingly the discrete time survival function is the same since $\theta(1) = 1 - S(1)$. The spells for Weibull with scale parameter 1 tend to be very short, and they start with a high hazard rate, so that there is very high probability of transition at the beginning of the spell. All the medians of the continuous time variables are in the first episode of their corresponded grouped time variables. It is likely that 63.21% of the transitions will be observed in the first episode. For the model with Weibull with shape 0.50 it is more likely to have more than half of the sample make transition out of the state before the end of the first half of episode 1. In Weibull variable with shape 0.25 it is more likely to have more than half of the sample make transition before the end of the first quarter of episode 1.

The simulations with Weibull variable with scale parameter $\lambda = 3$ is more related to economics. Baker and Melino (2000, page 365) use information about the duration of the unemployment spells and the hazard rate from Ham and Rea Jr (1987) to design the DGP of their simulation. They state that they "tried to generate hazards with the probability of exiting in the first period equal to the about 0.15 with about half the sample exiting by the fourth week". We find the simulation results with scale 3 are close to the hazard that they try to simulate. One can interpret the DGPs of a Weibull with scale parameter 3 and shape parameters 0.75 and 0.50, as simulations of unemployment spells by weeks with probability of exiting in the first week equal 0.3551 and 0.4386 respectively, and half of the individuals exit by the second week. Trying to make the probabilities closer to that in Baker and Melino (2000) means simulating longer spells and pay more computation cost in the grouped time PH estimators. In the model with $\lambda = 3$ and $k = 0.75$, about 95% of the individuals make a transition by the 13^{th} week (a little more than 3 months in the unemployment state) and 99% make transition by the 23^{th} week (after about 6 months in the unemployment state). We think that those are close figures to the common 'out of unemployment transition rates' in labour economics. Therefore we will show the simulation results of the DGP that are simulated from Weibull density function with scale 3 and shapes 0.75 and 0.50 in this chapter. The simulation results of all other Weibull parameter choices are presented in Appendix D.

3.5.2 The Hazard Estimators and the Measures of Performance in the Simulation Study

The Unconditional Hazard Models

The parametric proportional hazard PH models that are estimated for the unconditional hazard simulation are:

1. A complementary log-log model with Weibull baseline hazard specification. This model regress the pooled sample transition dummy on an intercept and $log(a_{ir})$ only. The model is denoted by θ^{WPH} and the estimated baseline hazard for each individual is the complementary log-log predicted probability at the last episode of that individual. To predict the hazard rate on episode j^o we use the log of the episode number on the right-hand-side of the estimated complementary log-log equation, as follows

$$\widehat{\theta}^{WPH}(j^o) = 1 - \exp\left\{-\exp\left[\widehat{\alpha}_0 + \widehat{\alpha}_1 \log(j^o)\right]\right\}. \tag{3.62}$$

2. A complementary log-log model with a step-function - piecewise linear- base-

Table 3.1: The expected value, the 50^{th}, 95^{th}, and the 99^{th} quantiles of the continuous Weibull transition variable and their location in the grouped time variable

Panel A: Continuous transition time

Weibull Coef.		$E[Y^c]$ and quantiles				Hazard	Survival function			
λ	k	μ_{Y^c}	$Q^{50}_{Y^c}$	$Q^{95}_{Y^c}$	$Q^{99}_{Y^c}$	$\phi_{Y^c}(1)$	$S_{Y^c}(1)$	$S_{Y^c}(Q^{50}_{Y^c})$	$S_{Y^c}(Q^{95}_{Y^c})$	$S_{Y^c}(Q^{99}_{Y^c})$
1	1.00	1	0.69	3.00	4.61	1	0.3679	0.5	0.05	0.01
	0.75	1.19	0.61	4.32	7.66	0.7500	0.3679	0.5	0.05	0.01
	0.50	2	0.48	8.97	21.21	0.5000	0.3679	0.5	0.05	0.01
	0.25	24	0.23	80.54	449.76	0.2500	0.3679	0.5	0.05	0.01
3	1.00	3	2.08	8.99	13.82	0.3333	0.7165	0.5	0.05	0.01
	0.75	3.57	1.84	12.96	22.99	0.3290	0.6449	0.5	0.05	0.01
	0.50	6	1.44	26.92	63.62	0.2887	0.5614	0.5	0.05	0.01
	0.25	72	0.69	241.62	1349.29	0.1900	0.4677	0.5	0.05	0.01

Panel B: Discrete transition time (grouped Y^c with interval width $b = 1$)

Weibull Coef.		Intervals of $E[Y^c]$ and the quantiles				Hazard	Discrete survival function			
λ	k	$j_{\{\mu_{Y^c}\}}$	$j_{\{Q^{50}_{Y^c}\}}$	$j_{\{Q^{95}_{Y^c}\}}$	$j_{\{Q^{99}_{Y^c}\}}$	$\theta(1)$	$S(1)$	$S\left(j_{\{Q^{50}_{Y^c}\}}\right)$	$S\left(j_{\{Q^{95}_{Y^c}\}}\right)$	$S\left(j_{\{Q^{99}_{Y^c}\}}\right)$
1	1	1	1	3	5	0.6321	0.3679	0.3679	0.0498	0.0067
	0.75	2	1	5	8	0.6321	0.3679	0.3679	0.0353	0.0086
	0.5	2	1	9	22	0.6321	0.3679	0.3679	0.0498	0.0092
	0.25	24	1	81	450	0.6321	0.3679	0.3679	0.0498	0.0100
3	1	3	3	9	14	0.2835	0.7165	0.0498	0.0001	0
	0.75	4	2	13	23	0.3551	0.6449	0.1860	0.0011	0
	0.5	6	2	27	64	0.4386	0.5614	0.2431	0.0061	0.0061
	0.25	72	1	242	1350	0.5323	0.4677	0.3679	0.0194	0.0026

[1] In Panel A: μ_{Y^c} is the expected value of Y^c, $Q^{\rho}_{Y^c}$ is the ρ^{th} quantile, ϕ_{Y^c} is the continuous time hazard rate at the index point of the first interval, $S_{Y^c}(Q^{\rho}_{Y^c})$ is the survival probability at the ρ^{th} quantile.

[2] Panel B: $j_{\{\mu_{Y^c}\}}$ is the episode that include the expected value of the continuous Weibull transition time, $j_{\{Q^{\rho}_{Y^c}\}}$ the episode that include the ρ^{th} quantile, $\theta(1)$ is the grouped time hazard rate in episode 1, $S\left(j_{\{Q^{\rho}_{Y^c}\}}\right)$ the survival probability at the end of the episode that include the ρ^{th} quantile of the continuous transition time.

[3] By construction: $S_{Y^c}(Q^{\rho}_{Y^c}) \geq S\left(j_{\{Q^{\rho}_{Y^c}\}}\right)$

line hazard. The episodes are grouped in the pooled sample to form the so-called a **series of episodes**. The distribution function of the grouped Weibull variable is considered in grouping the episodes, but the grouping is still somehow arbitrary. Let $d_{ir}^{(s)} = I\left(a_{ir} \in (\cdots]_s\right)$ be the dummy variable of the series of episodes s, where $s = 1, 2, \cdots, S$. Clearly the dummy variable of the first series is $d_{ir}^{(1)} = I\left(a_{ir} \in (1, \cdots]_1\right)$ and the last series is $d_{ir}^{(S)} = I\left(a_{ir} \in (\cdots, B]_S\right)$ with $S \leq B$. The pooled sample transition dummy is regressed on S series dummies, without including an intercept term, using the binary choice method with complementary log-log link function. The complementary log-log specification is then contain the step-function specification of the baseline hazard (also known as a semi-parametric baseline hazard). The step function PH model is denoted by θ^{SPH} in the simulation study, the first letter refers to **step-function**. To predict the hazard at episode j^o we use the estimated coefficient of the dummy variable of the series that include j^o in the complementary log-log equation. The hazard rate for the step s is predicted with the following equation

$$\widehat{\theta}^{SPH}(s) = 1 - \exp\left[-\exp\left(\sum_{s=1}^{S} \widehat{\alpha}_s d^{(s)}\right)\right], \tag{3.63}$$

where $\widehat{\alpha}_s$ is the estimated coefficient of the dummy variable corresponding to step s.

The coefficients are estimated using the Maximum Likelihood method of the discrete time PH models in Jenkins (2005, Chapter 6).

The nonparametric kernel estimators are :

1. The **density implied kernel hazard (DNK) estimators**. This estimators is the ratio between the estimated discrete time kernel density function and estimated discrete time kernel survival function. The estimator does not correct for the right censoring problem, where the research has not discuss any method of correcting the right censoring problem in the kernel density estimators. The kernel density estimator is obtained automatically since we are using the naïve bandwidths. The density implied kernel hazard estimators takes the form:

$$\widehat{\theta}^{DNK}(j^o) = \frac{\widehat{f}_{\widehat{\gamma}}(j^o)}{1 - \widehat{F}_{\widehat{\gamma}}(j^o)}, \tag{3.64}$$

where $\widehat{f}_{\widehat{\gamma}}(j^o) = \frac{1}{n}\sum_{i=1}^{n} l\left(J_i^o, j^o, \widehat{\gamma}\right)$ and $\widehat{F}_{\widehat{\gamma}}(j^o) = \sum_{r=1}^{j^o} \widehat{f}_{\widehat{\gamma}}(r),\ j^o = 1, 2, ..., \mathbf{B}$.

2. The grouped time external kernel hazard (EKH) estimator in Eq (3.47), $\widehat{\theta}^{EXK}(j^o)$.

The bandwidth $\widehat{\gamma}$ is estimated using the maximum likelihood cross-validation method MLCV using the unconditional density function in the density implied estimator, see Li and Racine (2003). The same bandwidths are used again in $\widehat{\theta}^{EKH}(j^o)$ as naïve bandwidths.

To measure the performance of the estimators in the simulation study, the distance measures from the true hazard are used. The sample mean square error and the sample Bias are measured is each Monte Carlo sample for all the estimators. The hazard is predicted in each sample for the episodes 1,2,$\cdots$,B', where $B' = J^o_{\{Q^{(99)}_{Yc}\}}$, the episode that include the 99^{th} quantile of the continuous Weibull variable, $Q^{(99)}_{Yc}$. For the samples that simulated from Weibull distribution with shape parameter 0.25, B' is chosen at 50 and 150 for the simulations with scale 1 and 3 respectively, which are equivalent to $B' = J^o_{\{Q^{(93)}_{Yc}\}}$. The predicted hazard rate of the parametric PH model with Weibull baseline hazard $\widehat{\theta}^{WPH}(j^o)$ are calculated for each episode in the range up to the episode B'. For the piecewise PH models, $\widehat{\theta}^{SPH}(s)$, the hazard rate are predicted for the step that include B'.

The sample mean squared error is computed in each Monte Carlo sample up to the interval $J^o_{\{Q^{(99)}_{Yc}\}}$. For sample q the predicted MSE is calculated as

$$\widehat{MSE}_q\left(\widehat{\theta}^{(\cdot)}\right) = \mathbf{S}^{-1}\sum_{r=1}^{S}\left(\theta(r) - \widehat{\theta}^{(\cdot)}_q(r)\right)^2, \quad q = 1, 2, ..., 500, \tag{3.65}$$

where $(\cdot)$ refers to the either SPH, DNK or EKH estimator.

Then the average of the MSE is taken over the Monte Carlo samples (aMSE) for each estimator as

$$\widehat{aMSE}^{(\cdot)} = 500^{-1}\sum_{q=1}^{500}\widehat{MSE}_q\left(\widehat{\theta}^{(\cdot)}\right). \tag{3.66}$$

The estimator that has the less $\widehat{aMSE}$ is the estimator that performed better in the study. Where less $\widehat{aMSE}$ means that the estimator predicts hazard rates that are closer to the true grouped time hazard in all episodes.

The absolute average Bias is computed in each sample also. For sample q the absolute bias is

$$\widehat{Bias}_q\left(\widehat{\theta}^{(\cdot)}\right) = \mathbf{S}^{-1}\sum_{r=1}^{S}\left|\theta(r) - \widehat{\theta}^{(\cdot)}_q(r)\right|, \quad q = 1, 2, ..., 500, \tag{3.67}$$

which is averaged over the Monte Carlo samples as

$$\widehat{Bias}^{(\cdot)} = 500^{-1}\sum_{q=1}^{500}\widehat{Bias}_q\left(\widehat{\theta}^{(\cdot)}\right). \tag{3.68}$$

For the WPH estimators $\widehat{MSE}_q\left(\widehat{\theta}^{(WPH)}\right)$ and $\widehat{Bias}_q\left(\widehat{\theta}^{(WPH)}\right)$ are computed by summing over the episodes up to B' instead of summing over the steps S.

To find the predictions of the kernel estimators in each step more work has been done. The kernel estimators are developed under the assumption of equal interval length. This is equivalent to a piecewise cloglog estimator with interval specific dummies. But in the piecewise cloglog model the predicted hazard on step (series of episodes) s is the average of the hazard rate of the episode in that step. Then the average of the predicted hazards for the episodes in the steps is taken.

The Conditional Hazard Models

The conditional hazard rate is estimated conditioning on two variables, a continuous independent variable which is simulated from $X^c \sim N(0,1)$ and a categorical independent variable which is defined as a binary variable with $Pr\left(X^d = 1\right) = 0.5$. The coefficients of those variables are chosen arbitrary after many investigations. The objective is to make the hazard conditioning on x^c significantly different between the groups of x^d, and to create a reasonable shift in the hazard rate when either variable is changed. So, we chose $\beta_{x^c} = 0.15$ and $\beta_{x^d} = 0.50$. This makes the proportional hazard increase by approximately 116% from the baseline hazard when $x^c = 1$ and drop to 80% of the baseline hazard when $x^c = -1$. On the other hand, the hazard rate of the group of individuals with $x^d = 1$ is 165% more than the hazard of the group in the baseline hazard. Those numbers are substantially lower than what was found by Baker and Melino (2000) and Nicoletti and Rondinelli (2010), where they simulate the same x^c but they set the coefficient equals 1, which means they assume a bigger contribution of the independent variable to the baseline hazard.

The contribution of the independent variables to the baseline hazard in the proportional hazard framework is specified as a change in the scale of the Weibull variable, see Jenkins (2005). Then the DGP of the simulation of the conditional hazard is designed as follows:

1. For the same choice of λ and k for the continuous Weibull transition time that is used in the unconditional hazard simulation, determine the proportion of censored observations π and the upper parameter of the Uniform distribution of the continuous censoring time, u_π. This is equivalent to setting the proportion of the censored observations in the baseline hazard.

2. Simulate X_i^c from $N(0,1)$ and X_i^d from Bernoulli with probability of success $Pr(X_i^d = 1) = 0.5$.

3. Simulated Y_i^c from $f_{Y^c|X_i^c,X_i^d}$. The mathematical proofs show that Y_i^c is simulated from the density $Y_i^c \sim Weibull(scale = \lambda exp\left[-\frac{1}{k}\left(0.15X_i^c - 0.5X_i^d\right)\right], shape = k)$. This means that we make the unconditional hazard in the first simulation the baseline hazard in the this simulation. Where for $X^c = 0$ and $X^d = 0$ the scale parameter reduces to λ.

4. Simulate C_i^c from r_{C^c}, the $Uniform(0, u_\pi)$ continuous density.

5. Compute the duration variable, $T_i^c = min(Y_i^c, C_i^c)$, and the right censoring indicator $\triangle_i = I(Y_i^c \leq C_i^c)$.

6. Bin T_i^c into equal intervals each of length $b = 1$.

7. Estimate the conditional hazard model using the estimators.

The parametric PH hazard estimators that are compared in the conditional hazard simulation study are a PH with a Weibull specification of the baseline hazard θ^{CWPH}, a PH with a step function specification of the baseline hazard θ^{CSPH}. The hazard rate is predicted at episode j^o conditional on chosen values of x^c and x^d, that are denoted by X^c and X^d. At episode j^o the hazard rate is predicted by the PH estimators as follows

1. The cloglog model with Weibull specification of the baseline hazard: find $log(j^o)$ and use the clolog link function with the estimated coefficients $\widehat{\alpha}_0$, $\widehat{\beta}_{x^c}$ and $\widehat{\beta}_{x^c}$, which are the coefficients of elapsed time in the state, the continuous independent variable and the discrete independent variable in the pooled sample respectively. The predicted hazard at episode j^o is given as follows

$$\widehat{\theta}^{CWPH}\left(j^o \left| X^c, X^d\right.\right) = 1 - \exp\left\{ - \exp\left[\widehat{\alpha}_0 \log\left(j^o\right) + \widehat{\beta}_{x^c} X^c + \widehat{\beta}_{x^d} X^d\right]\right\}. \tag{3.69}$$

2. The PH hazard with step function (piecewise) baseline hazard. To predict the hazard at episode j^o the estimator estimates the hazard for the step s that includes j^o. The hazard at each step s is predicted using the ML estimated coefficient of the dummy variable of s with the estimated coefficients of the continuous and discrete independent variables in the cloglog link function as follows:

$$\widehat{\theta}^{CSPH}\left(s \left| X^c, X^d\right.\right) = 1 - \exp\left\{ - \exp\left[\widehat{\alpha}_s + \widehat{\beta}_{x^c} X^c + \widehat{\beta}_{x^d} X^d\right]\right\}. \tag{3.70}$$

As in the unconditional PH parametric estimators, the coefficients of the transformed elapsed time variable and the coefficients of the independent variable are estimated using the ML method.

The kernel estimators of the conditional hazard are

1. The conditional density implied hazard estimator, $\widehat{\theta}^{CDNK}$, which is the estimated kernel conditional density function of the duration time divided by the estimated conditional survival function. The predicted hazard for episode j^o is as follows.

$$\widehat{\theta}^{CDNK}\left(j^o \left| X^c, X^d\right.\right) = \frac{\widehat{f}_{\widehat{\mathbf{h}}}\left(j^o \left| X^c, X^d\right.\right)}{1 - \widehat{F}_{\widehat{\mathbf{h}}}\left(j^o \left| X^c, X^d\right.\right)}. \tag{3.71}$$

where the survival function $\widehat{F}_{\widehat{\mathbf{h}}}(j^o \left| X^c, X^d\right.) = \sum_{r=1}^{j^o} \widehat{f}_{\widehat{\mathbf{h}}}(r \left| X^c, X^d\right.)$, $j^o = 1, 2, ..., \mathbf{B}$.

The conditional kernel density estimator $\widehat{\theta}^{CDNK}$ is automatically estimated since we are using the MLCV bandwidths of the conditional density as naïve bandwidths in the kernel hazard estimators.

2. Beran kernel hazard estimator, $\widehat{\theta}^{BKH}$, in Eq (4.29).
3. $\widehat{\theta}^{CEKH}$, the conditional external kernel hazard estimator in Eq (4.27).

The performance of the models is measured using 4 evaluation points of X^c and the categories of X^d. The values of X^c are: the mean of the Monte Carlo sample $\overline{X}^c_q$; the expected value of X^c, $\mu_{x^c} = 0$; μ_{x^c} plus 1 standard deviation, $\mu_{x^c} + \sigma_{x^c} = 1$; μ_{x^c} minus 1 standard deviation, $\mu_{x^c} - \sigma_{x^c} = -1$. The first two points are very close, the total number of the evaluation pairs **(evaluation points)** of X^c and X^d is 8 points. The evaluation point at $X^c = 0$ and $X^d = 0$ produces the baseline hazard. The hazard is predicted on each sample at each combination of X^c and X^d in all the intervals up to the interval of the 99^{th} quantile of continuous time $\left(Y^c|X^c = 0, X^d = 0\right)$.

To compute the measure of performance, the **average mean squared error (aMSE)**, for sample q, the predicted mean squared error at the evaluation point $\left(X^c, X^d\right)_e$, is computed by

$$\widehat{MSE}_{eq}(\widehat{\theta}^{(\cdot)}|\left(X^c, X^d\right)_e) = \mathbf{S}^{-1}\sum_{r=1}^{S}\left[\theta\left(r\left|\left(X^c, X^d\right)_e\right.\right) - \widehat{\theta}^{(\cdot)}_q\left(r\left|\left(X^c, X^d\right)_e\right.\right)\right]^2, \tag{3.72}$$

where $e = 1, 2, \cdots, 8$ denotes the evaluation point and $q = 1, 2, \cdots, 500$. $(\cdot)$ refers to the estimators CSPH, BKH, CEKH, CDNK. For the CWPH the summation is over the intervals, B', not over the steps S. The average mean squares error for each estimator $\widehat{aMSE}^{(\cdot)}$ is calculated by taking the average over all samples and all the evaluation points.

$$\widehat{aMSE}^{(\cdot)} = 500^{-1}\sum_{q=1}^{500}\left(8^{-1}\sum_{e=1}^{8}\widehat{MSE}_{eq}\left(\widehat{\theta}^{(\cdot)}\right)\right). \tag{3.73}$$

On the other hand, the absolute bias for sample q at the evaluation point e is computed as

$$\widehat{Bias}_q(\widehat{\theta}^{(\cdot)}|\left(X^c, X^d\right)_e) = \mathbf{S}^{-1}\sum_{r=1}^{S}|\theta\left(r\left|\left(X^c, X^d\right)_e\right.\right) - \widehat{\theta}^{(\cdot)}_q\left(r\left|\left(X^c, X^d\right)_e\right.\right)|, \tag{3.74}$$

and the **average absolute bias** for estimator $(\cdot)$ in all samples at all the evaluation point is

$$\widehat{Bias}^{(\cdot)} = 500^{-1}\sum_{q=1}^{500}\left(8^{-1}\sum_{e=1}^{8}\widehat{Bias}_{eq}\left(\widehat{\theta}^{(\cdot)}\right)\right). \tag{3.75}$$

3.5.3 The Simulations Results

The simulation results are presented in two parts: the first part - Simulation Results Part A - is related to the parametric estimators only, the second part - Simulation Results Part B - include the comparison between the estimators, the parametric PH and the KH estimators. The discussion in the first part focuses on the Monte Carlo estimated coefficients of the AFT and PH models only. After we show that the coefficients of the parametric models are estimated correctly in the first part, we proceed into the comparison of the estimators using the performance measures in the second part. The comparison in the second part include all the discrete time hazard estimators in the simulation study, the AFT estimator is not included in the performance measures.

Simulation Results Part A: Comparison between the estimated coefficients of the parametric estimators

The scale and shape parameters of Weibull continuous time are estimated directly in the continuous time hazard regression, the estimators that are denoted by AFT (accelerated failure time) and CAFT (conditional accelerated failure time) in the unconditional and conditional DGP results respectively. For WPH and CWPH estimators the parameters of the continuous density function of the underlying Weibull variable are not identified without more restrictions[8]. The coefficients of the independent variables that are estimated by the CAFT estimator are denoted by $\widehat{\beta}^{*AFT}_{x^c}$ and $\widehat{\beta}^{*AFT}_{x^d}$ for x^c and x^d respectively in Tables 3.2 and 3.3. The same coefficients in the PH form are denoted by $\widehat{\beta}^{AFT}_{x^c}$ and $\widehat{\beta}^{AFT}_{x^d}$ respectively. The transformation of the coefficients from CAFT to PH form depends on the Weibull density shape parameter as follows, $\boldsymbol{\beta}^{*^{AFT}} = -k^{-1}\boldsymbol{\beta}^{AFT}$.

The estimated values of the scale and shape parameters by the AFT and CAFT estimators, show that the Weibull continuous time hazard model is estimated correctly. This implies that the continuous time PH model could follow directly. However, the continuous time PH is not included. We will focus on the estimated coefficients, where they are sufficient to examine the design of the DGP and the grouping scheme. The estimated coefficients for the independent variables that shown in the tables are almost the same for the continuous time estimator and the discrete time estimators. The coefficient of the continuous independent variable is set at $\beta_{x^c} = 0.15$ and of the dummy independent variable is set at $\beta_{x^c} = 0.50$. Both values are estimated correctly by the continuous time CAFT estimator and the discrete time CWPH and CSPH estimators , see the tables in Appendix D.

Hence, the effect of the independent variables to the hazard function is correctly estimated in the continuous time model and in the grouped the time models. Which means that the observed heterogeneity function is estimated correctly by the PH parametric estimators. The coefficients that are estimated in the continuous time estimator seem

[8]One of those restrictions is to restrict the values the scale and the shape to 1 as shown in Appendix D

slightly better than the coefficients that are estimated in the grouped time estimators. On the other hand, the coefficients that are estimated in the fully parametric grouped time model, CWPH, are also slightly better than the coefficients that are estimated in the step-function grouped time PH model, CSPH. This is an intuitive result since the true model is simulated from a known Weibull conditional hazard model.

The PH models estimate the baseline hazard and the observed heterogeneity functions jointly. The intercept and the coefficient of $log(j^o)$ in the WPH and CWPH models and the coefficients of the steps dummies in the SPH and CSPH models are the coefficient that estimate the baseline hazard function. The estimated coefficient of the independent variables are the coefficient that estimate the observed heterogeneity function. If the Weibull shape and scale parameters are identified in the model, they become predictable by the intercept and the coefficient of $log(j^o)$ in the WPH and CWPH models. As the mathematical proofs and the simulation results in Appendix D show, the shape parameters is identified in our simulations when $k = 1$ only, on the Exponential case, where the shape parameter is $k = 1 - \alpha_{log(j^o)}$. In the simulations $1 - \widehat{\alpha}_{log(j^o)}$ is used to find the estimated shape parameter. The identification of the parameters of Weibull continuous variable in the PH models is not important in our simulation study, the baseline hazard is estimated correctly for all values of k. The identification of the Weibull parameters in the PH is not concerned in the performance measures. We argue that if we tried to include assumptions to make the Weibull parameters identified to all WPH and CWPH models, this may restrict the comparison and affect our research objectives.

Simulation Results Part B: Comparison between discrete time hazard estimators

The nonparametic kernel estimators do not produce coefficients, the only estimated parameters are the bandwidths. So, the comparison between parametric and nonparametric kernel estimates is slightly difficult. The researcher attempt to use a method that makes the reading of the simulation results and the comparison between the estimators easy, with this large number of hazard estimators that are compared. The estimators that are included in the simulation study are compared using distance measures between the true hazard and the predicted hazard that estimated by each PH or KH estimator. The estimator that predicts a hazard rate that closer to the true hazard is better than the other estimators. Counting the number of (estimators$\times$baseline hazards$\times$samples)[9]. We estimate each hazard estimator a number of times that equals the to the product of the chosen Weibull shape parameters, k, by the number of the chosen Weibull scale parameters, λ, (baseline hazards). This is repeated at each chosen sample size n and proportions of the right censored observation, π (samples). The result of the overall number of times that each estimator is estimated is 160. The simulation study compares between 4 unconditional hazard estimators and 5 conditional hazard

[9]The proportion of right censored observations and the sample size are the two characteristics that define each sample in the simulation study.

Table 3.2: Monte Carlo estimated coefficients of the of the conditional PH models for the Weibull with scale=3 and shape=0.75 simulated baseline hazard models.

	n	$\widehat{\pi}$	$\widehat{\lambda}^{AFT}$	$\widehat{k}^{AFT}$	$\widehat{\beta}^{*AFT}_{x^c}$	$\widehat{\beta}^{*AFT}_{x^d}$	$\widehat{\beta}^{AFT}_{x^c}$	$\widehat{\beta}^{AFT}_{x^d}$	$\widehat{\alpha}^{AFT}_{log(j)}$	$\widehat{\beta}^{CWPH}_{x^c}$	$\widehat{\beta}^{CWPH}_{x^d}$	$\widehat{\beta}^{CSPH}_{x^c}$	$\widehat{\beta}^{CSPH}_{x^d}$	$\widehat{\gamma}_0$	$\widehat{h}_{x^c}$	$\widehat{\gamma}_{x^d}$
1	100	0%	3.0004	0.7751	-0.1972	-0.6704	0.1520	0.5192	-0.8828	0.1525	0.5206	0.1543	0.5288	0.289853	2.545958	0.246827
2	200	0%	3.0018	0.7564	-0.1955	-0.6609	0.1479	0.4999	-0.8684	0.1469	0.4987	0.1493	0.5065	0.202582	2.252689	0.224032
3	500	0%	3.0007	0.7518	-0.2081	-0.6712	0.1565	0.5047	-0.8617	0.1552	0.5005	0.1573	0.5071	0.113578	1.737428	0.166733
4	1000	0%	3.0119	0.7523	-0.1984	-0.6707	0.1493	0.5046	-0.8631	0.1477	0.4994	0.1496	0.5056	0.069607	1.557868	0.119116
5	100	2.218%	3.0269	0.7685	-0.2122	-0.6752	0.1627	0.5201	-0.8902	0.1641	0.5198	0.1668	0.5302	0.288871	2.004303	0.264756
6	200	2.134%	3.0185	0.7598	-0.1975	-0.6696	0.1501	0.5087	-0.8792	0.1499	0.5082	0.1516	0.5149	0.189110	2.325629	0.220240
7	500	2.179%	3.0193	0.7505	-0.2019	-0.6710	0.1516	0.5035	-0.8716	0.1504	0.5005	0.1527	0.5063	0.110090	1.820908	0.171000
8	1000	2.114%	3.0059	0.7501	-0.2021	-0.6701	0.1515	0.5027	-0.8617	0.1504	0.4969	0.1521	0.5026	0.065707	1.528744	0.124665
9	100	5.488%	2.9994	0.7672	-0.2071	-0.6452	0.1594	0.4936	-0.8767	0.1588	0.4938	0.1603	0.5007	0.258930	2.020358	0.286649
10	200	5.535%	3.0564	0.7588	-0.2047	-0.6912	0.1551	0.5242	-0.8907	0.1550	0.5227	0.1566	0.5300	0.169470	2.201203	0.223013
11	500	5.473%	3.0117	0.7547	-0.2072	-0.6688	0.1563	0.5048	-0.8741	0.1554	0.5018	0.1570	0.5072	0.089128	1.805558	0.173884
12	1000	5.362%	2.9902	0.7522	-0.1995	-0.6595	0.1500	0.4961	-0.8652	0.1487	0.4924	0.1504	0.4973	0.047267	1.589593	0.134860
13	100	10%	3.0375	0.7732	-0.1960	-0.6587	0.1517	0.5091	-0.9075	0.1545	0.5086	0.1566	0.5176	0.217274	2.199813	0.295689
14	200	9.875%	3.0406	0.7584	-0.1955	-0.6821	0.1481	0.5173	-0.8891	0.1475	0.5144	0.1491	0.5202	0.127827	2.487766	0.239130
15	500	9.815%	2.9822	0.7527	-0.1979	-0.6566	0.1490	0.4940	-0.8658	0.1481	0.4889	0.1493	0.4935	0.057824	1.873160	0.192550
16	1000	9.994%	2.9994	0.7506	-0.1982	-0.6666	0.1487	0.5003	-0.8718	0.1475	0.4963	0.1489	0.5006	0.025453	1.586350	0.138000
17	100	25.9%	3.1131	0.7672	-0.2126	-0.7005	0.1621	0.5352	-0.9204	0.1619	0.5288	0.1624	0.5320	0.071853	2.617331	0.342571
18	200	25.59%	3.0411	0.7587	-0.2043	-0.6682	0.1546	0.5072	-0.9045	0.1533	0.5010	0.1536	0.5034	0.023082	2.465041	0.281123
19	500	25.72%	3.0188	0.7529	-0.2011	-0.6666	0.1515	0.5014	-0.8901	0.1503	0.4956	0.1506	0.4967	1.6×10^{-8}	1.830854	0.187128
20	1000	25.71%	3.0138	0.7502	-0.1987	-0.6749	0.1491	0.5061	-0.8900	0.1474	0.5002	0.1478	0.5014	9.5×10^{-9}	1.400444	0.127707

[1] The baseline hazard is simulated from Weibull with scale $\lambda = 3$ and shape $k = 0.75$.
[2] The conditional transition time for the proportional hazard is simulated from the function $Y_i^c \sim Weibull(scale = \lambda exp\left[-\frac{1}{k}\left(0.15X_i^c - 0.5X_i^d\right)\right], shape = k)$. The interval length $b = 1$. The censoring time is simulated from $C_i^c \sim Uniform(0, u_\pi)$, where π is the proportion of the right censored observations.
[3] The distribution of the covariates are $X_i^c \sim Normal(0,1)$ and $Pr(X_i^d = 1) = 0.5$.
[4] The average Mean Square Error aMSE of the models in this table are shown in Table D.29 below.

Table 3.3: Monte Carlo estimated coefficients of the of the conditional PH models for the Weibull with scale=3 and shape=0.50 simulated baseline hazard models.

	n	$\widehat{\pi}$	$\widehat{\lambda}^{AFT}$	$\widehat{k}^{AFT}$	$\widehat{\beta}^{*AFT}_{x^c}$	$\widehat{\beta}^{*AFT}_{x^d}$	$\widehat{\beta}^{AFT}_{x^c}$	$\widehat{\beta}^{AFT}_{x^d}$	$\widehat{\alpha}^{AFT}_{log(j)}$	$\widehat{\beta}^{CWPH}_{x^c}$	$\widehat{\beta}^{CWPH}_{x^d}$	$\widehat{\beta}^{CSPH}_{x^c}$	$\widehat{\beta}^{CSPH}_{x^d}$	$\widehat{\gamma}_0$	$\widehat{h}_{x^c}$	$\widehat{\gamma}_{x^d}$
1	100	0%	3.1550	0.5127	-0.2968	-1.0220	0.1518	0.5242	-0.6754	0.1542	0.5223	0.1565	0.5309	0.468648	2.212406	0.249877
2	200	0%	3.0489	0.5078	-0.2839	-0.9983	0.1442	0.5064	-0.6569	0.1445	0.5036	0.1476	0.5109	0.389423	2.547175	0.215248
3	500	0%	3.0033	0.5031	-0.2982	-0.9863	0.1500	0.4964	-0.6505	0.1481	0.4931	0.1510	0.5008	0.287831	1.898686	0.185768
4	1000	0%	3.0117	0.5018	-0.3010	-1.0032	0.1510	0.5034	-0.6518	0.1488	0.4989	0.1519	0.5072	0.209793	1.742646	0.154272
5	100	2.468%	3.0705	0.5148	-0.3086	-0.9940	0.1589	0.5123	-0.6602	0.1592	0.5176	0.1616	0.5246	0.432247	2.277321	0.267749
6	200	2.339%	3.0259	0.5062	-0.3028	-1.0066	0.1535	0.5094	-0.6490	0.1532	0.5064	0.1554	0.5132	0.344757	2.200912	0.229283
7	500	2.379%	3.0016	0.5031	-0.3040	-0.9990	0.1529	0.5026	-0.6394	0.1514	0.4970	0.1534	0.5048	0.234231	2.087896	0.198443
8	1000	2.37%	3.0058	0.5015	-0.2979	-1.0006	0.1494	0.5018	-0.6426	0.1480	0.4970	0.1503	0.5037	0.149357	1.812612	0.168394
9	100	5.866%	3.1239	0.5123	-0.3058	-0.9966	0.1563	0.5113	-0.6558	0.1562	0.5124	0.1583	0.5213	0.371224	2.511619	0.277746
10	200	5.651%	3.0624	0.5056	-0.3006	-1.0087	0.1517	0.5100	-0.6432	0.1521	0.5080	0.1538	0.5142	0.268259	2.375836	0.258794
11	500	5.676%	2.9917	0.5016	-0.3055	-0.9987	0.1531	0.5007	-0.6313	0.1526	0.4973	0.1540	0.5025	0.142730	2.296658	0.226942
12	1000	5.748%	3.0232	0.5017	-0.2990	-1.0054	0.1500	0.5043	-0.6377	0.1491	0.4999	0.1507	0.5054	0.079707	1.916980	0.176252
13	100	10.05%	3.0304	0.5075	-0.3092	-0.9810	0.1567	0.4969	-0.6339	0.1571	0.5022	0.1589	0.5075	0.287587	2.768975	0.330533
14	200	9.927%	3.0385	0.5078	-0.2996	-1.0034	0.1522	0.5093	-0.6346	0.1524	0.5078	0.1535	0.5141	0.180688	2.432865	0.283399
15	500	9.94%	3.0296	0.5017	-0.2962	-1.0120	0.1485	0.5076	-0.6283	0.1479	0.5053	0.1487	0.5090	0.073304	2.317099	0.233370
16	1000	10.04%	3.0141	0.4999	-0.2992	-1.0009	0.1495	0.5002	-0.6253	0.1487	0.4971	0.1495	0.5002	0.030535	1.946737	0.186236
17	100	25.53%	3.1718	0.5153	-0.3217	-1.0234	0.1652	0.5231	-0.6211	0.1654	0.5138	0.1651	0.5151	0.060398	2.904028	0.364193
18	200	25.48%	3.0736	0.5055	-0.3083	-0.9875	0.1552	0.4975	-0.6204	0.1554	0.4958	0.1558	0.4965	0.000000	2.862604	0.313535
19	500	25.63%	3.0548	0.5025	-0.3092	-1.0070	0.1553	0.5056	-0.6199	0.1537	0.5009	0.1536	0.5017	2.5×10^{-9}	1.895395	0.194318
20	1000	25.51%	3.0144	0.5010	-0.3070	-1.0068	0.1536	0.5042	-0.6164	0.1525	0.4998	0.1526	0.5003	2.5×10^{-9}	1.419826	0.128635

[1] The baseline hazard is simulated from Weibull with scale $\lambda = 3$ and shape $k = 0.50$.
[2] The conditional transition time for the proportional hazard is simulated from the function $Y_i^c \sim Weibull(scale = \lambda exp\left[-\frac{1}{k}\left(0.15X_i^c - 0.5X_i^d\right)\right], shape = k)$. The interval length $b = 1$. The censoring time is simulated from $C_i^c \sim Uniform(0, u_\pi)$, where π is the proportion of the right censored observations.
[3] The distribution of the covariates are $X_i^c \sim Normal(0,1)$ and $Pr(X_i^d = 1) = 0.5$.
[4] The average Mean Square Error aMSE of the models in this table are shown in Table D.30 below.

estimators. This gives a total of 640 unconditional and 800 conditional Monte Carlo predicted hazard, each repeated 500 time.

We aim to provide sufficient information about the results of the simulation study. So, we follow a method that uses a graphical presentation to a rescaled aMSE measure. For each estimator on each (baseline hazard $\times$ sample) the rescaled distance measure, $log(aMSE \times 10^5)$, is represented by a marker in Figures 3.2 to 3.5 below. The transformation of the aMSE to $log(aMSE \times 10^5)$ makes the comparison easier, Figures 3.2 and 3.3 present the results of the unconditional hazard simulation, Figures 3.4 and 3.5 present the results of the conditional hazard simulation. The $log(aMSE \times 10^5)$ of the conditional estimators in Tables 3.2 and 3.3 are shown in the middle columns in Figure 3.5. The $log(aMSE \times 10^5)$ of the PH with Weibull baseline hazard estimator, WPH and CWPH estimators, is shown by black markers. The external kernel hazard estimators are shown by blue markers in all plots.

Figures 3.2 and 3.3 show that (C)WPH model performance deteriorates dramatically when the shape parameter decrease. $log(aMSE \times 10^5)$ of the (C)WPH model is higher than that for other model. At low value of the shape parameter, $k = 0.25$, the conditional hazard of the grouped time model takes the form:

$$\theta\left(j^o|\mathbf{x}\right) = 1 - exp\left[-exp\left(klog\left(\frac{1}{\lambda}\right) + log\left({j^o}^k - \left(j^o - 1\right)^{k-1}\right) + \mathbf{x}^T\boldsymbol{\beta}\right)\right] \qquad (3.76)$$

Identifying the shape parameter of the underlying continuous time becomes difficult in the model. (C)WPH model use the term $\alpha\log(j^o)$ in the cloglog link function in an attempt to estimate the cumulative hazard the takes the form $klog\left(\frac{1}{\lambda}\right) + log\left({j^o}^k - \left(j^o - 1\right)^{k-1}\right)$. α is assumed to be an estimator of $k-1$, but it is identified with respect to k only when $k = 1$ where the coefficient α estimate is 0. When k is different than 1 k becomes unidentified to α, with k goes to a small number the identification becomes more difficult. Then the grouped time model cannot identify the parameters of the continuous time model, see Jenkins (2005) for detailed discussion.

The (C)WPH perform the worst on the models that are simulated from Weibull with shape parameter $k = 0.25$. The (C)WPH model defines the baseline function as $1 - exp(-exp(\alpha_0 + \alpha_1 log j^o))$ whilst the function takes the form $1 - exp[-exp(klog(\frac{1}{\lambda}) + log((j^o)^k - (j^o - 1)^k))]$. At $k = 1$ the true baseline hazard reduces to $1 - exp(-exp(log(\frac{1}{\lambda})))$. The intercept α_0 is an estimate of the $log(\frac{1}{\lambda})$ and the coefficient $\alpha_1 = 1 - k = 0$. Then the (C)WPH capture the parameters of Weibull variables in the model. When $k \neq 1$ the (C)WPH model fails to identify the parameters of the Weibull underlying transition time. With the k becomes largely different that 1, the difference between (C)WPH model definition of the baseline hazard as the true baseline hazard becomes large. The limits $\underset{k \longrightarrow c}{lim}(1 - k)log j^o$ and $\underset{k \longrightarrow c}{lim} log((j^o)^k - (j^o - 1)^k)$ are equal at $k = 1$ only, which explains the reason that the performance of the (C)WPH model deteriorates substantially when the shape parameters reduced.

The semi-parametric and the nonparametric estimation techniques generally need large sample size to preform good, see Li and Racine (2007). This explains the subitantial increase in the performance of the distance measure in the large sample size, $n = 500$ and $n = 1000$, in the plots in Figures 3.2 to 3.5 below. The aMSE is negatively related with the sample size, the downward slope of the $log(aMSE \times 10^5)$ dotted line in the plots is resulted from the increase in the sample size. In contrast to Weibull baseline hazard estimators, the semi-parametric and the nonparametric estimators, as implied by the plots, are not affected by the shape unmatchability between the discretised time density function and the continuous time density function.

The aMSE for the unconditional hazard estimators is very close to the aMSE of the conditional hazard at the baseline hazard evaluation points. This is an advantage in the design of the DGP, where the proportional effect of the independent variables is simulated upon the unconditional hazard. The distance between the true conditional hazard and the hazard that estimated by any of the estimators is consistent. The quality of the hazard predictions are similar for the estimators among all the evaluation points, except for the density implied hazard estimator.

The proportional hazard estimators are generally better than the kernel estimators. The performance of the kernel estimators deteriorate substantially when the hazard is predicted at the evaluation points with $X^d = 1$. Kernel estimators failed to capture the proportional contribution of the categorical independent variable in the simulation study. In Figures 3.6 to Figures 3.11 the conditional hazard is predicted at the baseline hazard evaluation points, $\left(X^c = 0, X^d = 0\right)$, in addition to the points $\left(X^c = 0, X^d = 1\right)$ and $\left(X^c = 1, X^d = 0\right)$ and $\left(X^c = 1, X^d = 1\right)$, for a sample size 500 and proportion of right censored observations 0%, 5% and 25%. The right column shows the true hazard (the black curve) and the PH estimators predictions, the left column shows the same true hazard and the predictions of the kernel estimators. It can be seen in the charts that how the PH estimators produce estimates that are very close to the true hazard at all the evaluation points. The upward shift in the conditional hazard occurs when the dummy independent variable increases from 0 to 1. The PH estimators captured the upward shift better than the kernel estimators. Where the kernel estimators have underestimated the contribution of the categorical independent variable as shown in the plots.

The right censoring problem causes a reduction in the observed duration in the state in the simulation study. When the Uniform density is higher than the Weibull density the probability of the right censoring is higher than the probability of transition. This occurs for a short period at the beginning of the spell, but the probability of transition increases as the Weibull density function increases upward and exceeds the height of the Uniform density function. At the end of the spell the probability of transition decreases, the Uniform density function becomes higher than the Weibull density function, this leads to a drop in the probability of transition and an increase in the probability of the right censoring. The drop in the duration of the spell is clearer in the samples that are simulated from Weibull(scale=3, shape=0.50), where the spell length is dropped from 22 episodes to almost the half, to about 10 episodes only. The right censoring problem seems to make the estimators underestimate the true hazard rate at the end of the spells, but the underestimation is higher in the kernel estimators when the hazard is predicted at $X^d = 1$.

Lastly, the average bias at the first episode is calculated for the conditional hazard at the 8 evaluation points of X^c and X^d. The average of the absolute bias of the predicted hazard rate for the episodes up to the episode that include the 99^{th} quantile of the continuous time is computed for each estimator, with the percentage bias at the first episode. The bias in the first episode can be an indicator of the Boundary Bias problem, if the problem still in the smoothing of the grouped time variable. Where the predicted hazard on the first episode will have high percentage bias if the Boundary Bias still exist.

The result of the bias measures for the duration models that are simulated from Weibull with shape parameters 0.75 and 0.50 and scale parameter 3, are shown in Tables 3.6 and 3.7. The percentage bias of the first episode is higher in the PH with Weibull baseline estimator than in the step-function PH estimator. The first episode bias in the kernel hazard estimator, in contrast to the PH estimator, increases with the increase in the sample size and is not affected much by the proportion of the right censored observations, except in the density implied estimator because it does not control for the right censoring problem. The first episode bias is substantially lower than half, which is the amount of Boundary Bias at the continuous time kernel hazard models. However, this result is just implied from the simulation results, we haven't studied the theoretical properties of the discrete time KH estimators. Part of this bias, however, is due to the use of inefficient bandwidths estimates in the discrete time kernel hazard estimators. The absolute bias in the lower panel in Tables 3.6 and 3.7 shows the performance of the estimators in all episodes. The estimators with the lowest absolute bias is the WPH and CWPH followed by SPH and CSPH estimators. This illustrates that the bias in the kernel predictions is high.

The results of the distance measures and that shown in the hazard plots, in addition to the absolute bias results, imply that the kernel estimators have not performed well in the simulation study. As all kernel estimators, the choice of the method to estimate the bandwidths in the model is the corner stone of the good results. The choice of using the naive bandwidths in this simulation study lead to estimate poor results by the kernel hazard estimators. The naive bandwidths are not estimated by trading between the variance and the bias in the hazard formula. Where this might likely be the reason that the kernel hazard estimators failed to capture the effect of the categorical variable in the simulation study.

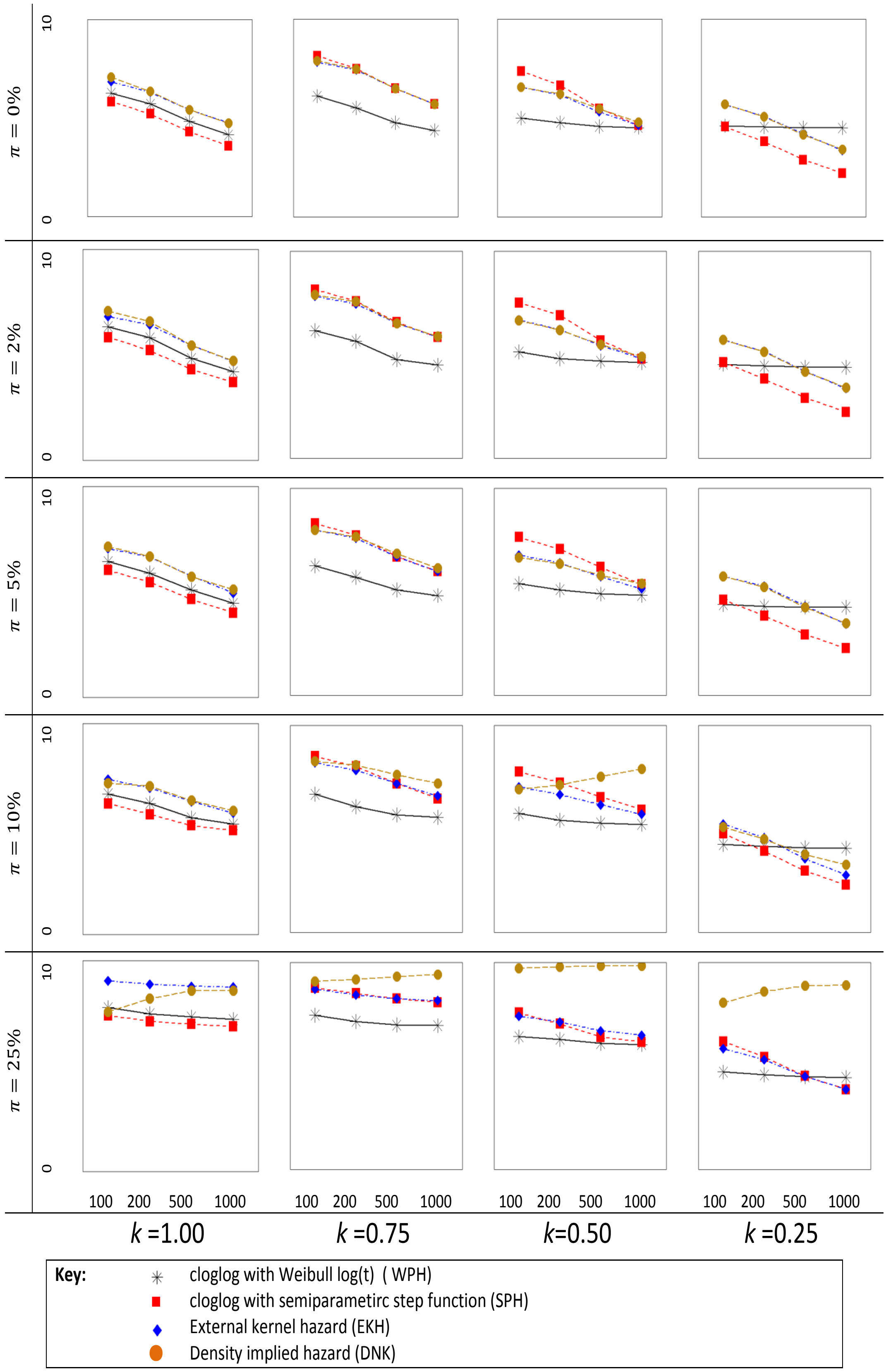


Figure 3.2: $log(aMSE \times 10^5)$ for the simulations of the unconditional hazard with scale parameter 1.

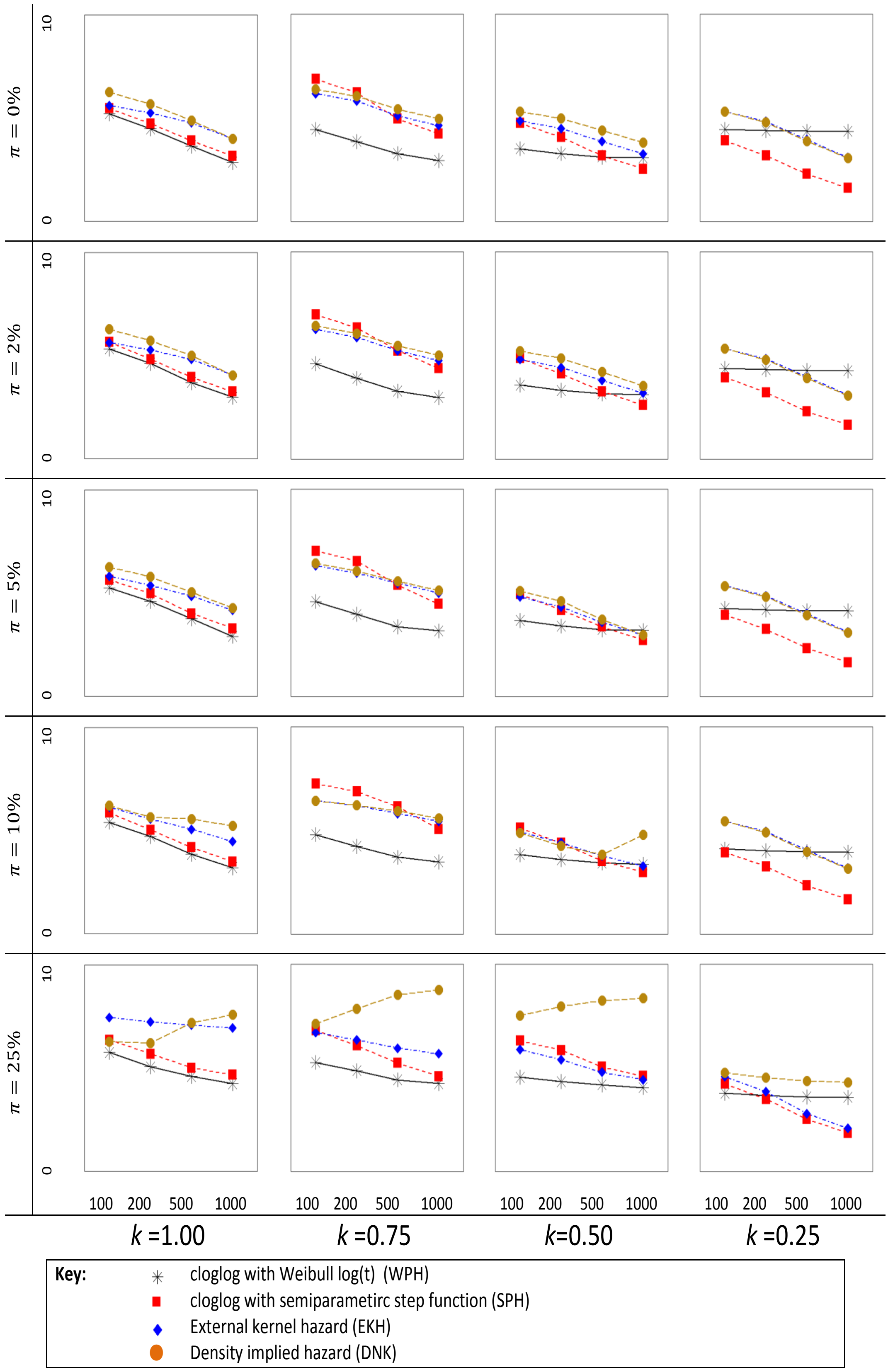


Figure 3.3: $log(aMSE \times 10^5)$ for the simulations of the unconditional hazard with scale parameter 3.

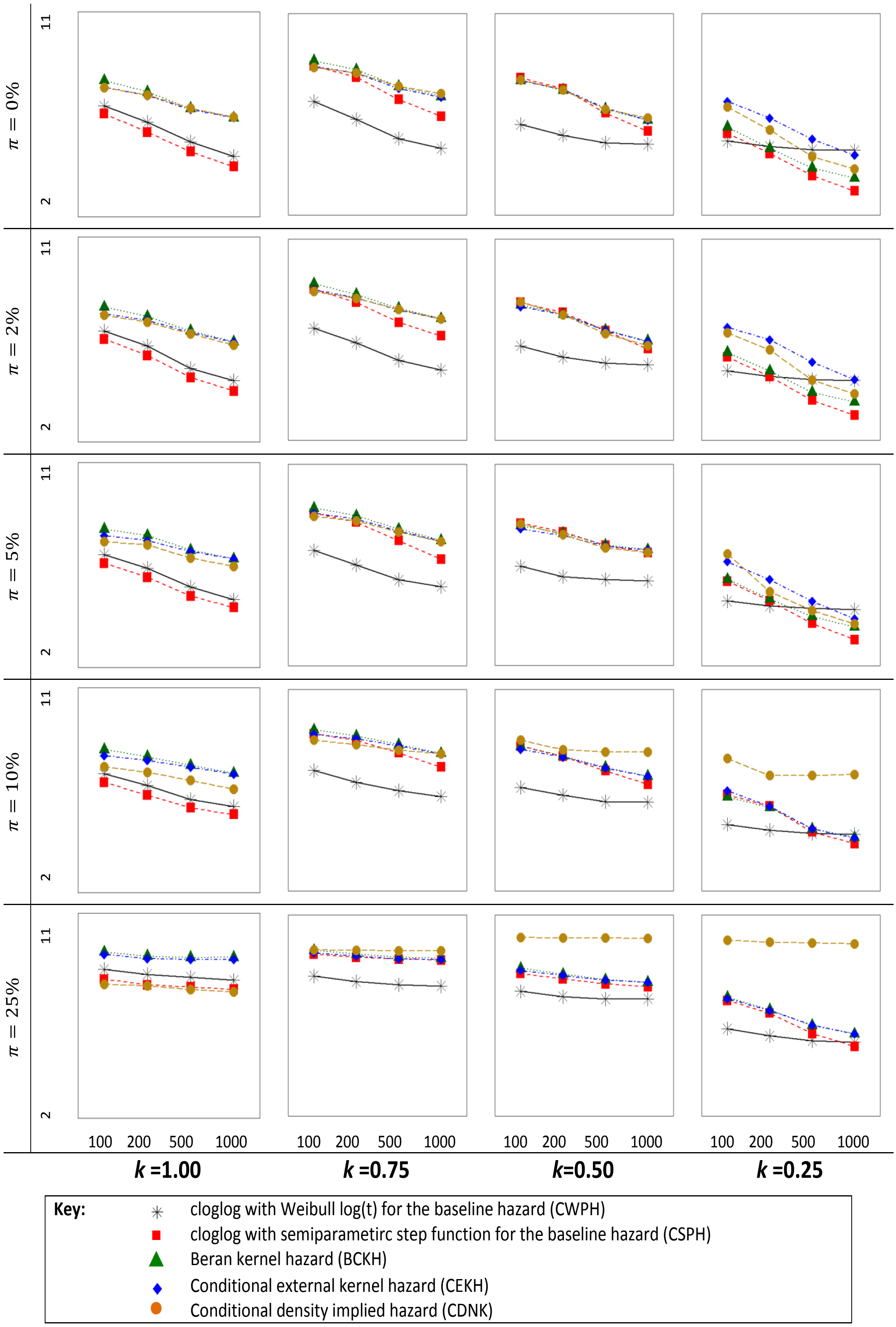


Figure 3.4: $log(aMSE \times 10^5)$ for the simulations of the conditional hazard with scale parameter 1.

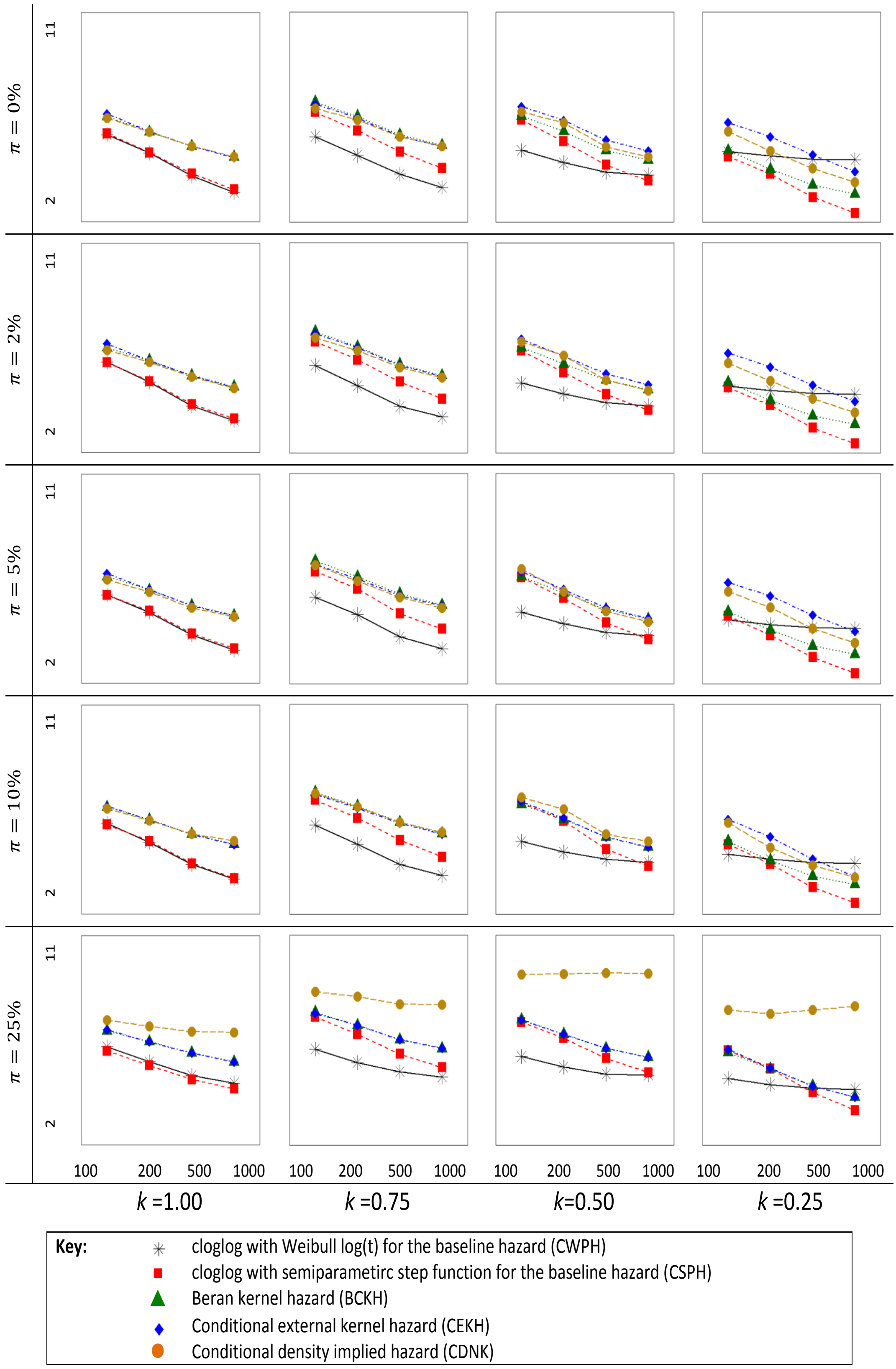


Figure 3.5: $log(aMSE \times 10^5)$ for the simulations of the conditional hazard with scale parameter 3.

Table 3.4: aMSE of the conditional hazard models with the baseline hazard that follows Weibull with scale=3 and shape=0.75.

		aMSE of the baseline hazard				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.001694	0.005676	0.008143	0.006798	0.005532
200	0%	0.000803	0.002355	0.003637	0.003131	0.002778
500	0%	0.000319	0.000967	0.001554	0.001361	0.001280
1000	0%	0.000206	0.000477	0.000807	0.000743	0.000714
100	2.2%	0.001541	0.005497	0.008132	0.007031	0.005891
200	2.1%	0.000737	0.002682	0.003989	0.003395	0.003334
500	2.2%	0.000311	0.000947	0.001519	0.001365	0.001456
1000	2.1%	0.000193	0.000468	0.000797	0.000728	0.000817
100	5.5%	0.001711	0.005732	0.008745	0.007172	0.007256
200	5.5%	0.000718	0.002484	0.003811	0.003250	0.003772
500	5.5%	0.000333	0.000990	0.001563	0.001408	0.002060
1000	5.4%	0.000210	0.000502	0.000823	0.000756	0.001329
100	10%	0.001793	0.006454	0.009105	0.007630	0.009758
200	9.9%	0.000801	0.002887	0.004133	0.003642	0.005811
500	9.8%	0.000362	0.001158	0.001765	0.001617	0.003623
1000	9.4%	0.000232	0.000553	0.000847	0.000803	0.002485
100	25.9%	0.002728	0.011895	0.016292	0.014843	0.059273
200	25.6%	0.001444	0.004881	0.006624	0.006343	0.047432
500	25.7%	0.001094	0.002149	0.002906	0.002885	0.040385
1000	25.7%	0.000903	0.001331	0.001744	0.001741	0.037771

		aMSE of the conditional hazard at 8 combination of values for x^c and x^d				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.003341	0.009332	0.013343	0.012625	0.010603
200	0%	0.001373	0.003686	0.006712	0.006148	0.005648
500	0%	0.000555	0.001550	0.003304	0.003113	0.003030
1000	0%	0.000350	0.000764	0.001979	0.001916	0.001895
100	2.2%	0.003203	0.009145	0.013748	0.012460	0.011264
200	2.1%	0.001445	0.004348	0.006917	0.006440	0.005787
500	2.2%	0.000555	0.001516	0.003299	0.003129	0.002915
1000	2.1%	0.000336	0.000761	0.002096	0.002020	0.001866
100	5.5%	0.003175	0.009109	0.013774	0.012160	0.011140
200	5.5%	0.001303	0.004186	0.007372	0.006537	0.006503
500	5.5%	0.000573	0.001586	0.003488	0.003312	0.003083
1000	5.4%	0.000349	0.000791	0.002191	0.002109	0.001971
100	10%	0.003217	0.010190	0.014259	0.012421	0.013020
200	9.9%	0.001459	0.004631	0.007407	0.006826	0.007012
500	9.8%	0.000621	0.001783	0.003923	0.003750	0.003860
1000	9.4%	0.000390	0.000893	0.002436	0.002374	0.002499
100	25.9%	0.004646	0.018034	0.022379	0.020905	0.050521
200	25.6%	0.002571	0.007859	0.012088	0.011810	0.038542
500	25.7%	0.001821	0.003502	0.006851	0.006828	0.032688
1000	25.7%	0.001433	0.002094	0.004582	0.004578	0.030969

[1] The point of (X^c, X^d) that used to calculated the average aMSE are $(0,0), (1,0), (-1,0), (\overline{X}^c, 0), (0,1), (1,1), (-1,1), (\overline{X}^c, 1)$. The first point measures the hazard of the baseline group.

[2] The interval length is $b = 1$ and the number of intervals that used in the aMSE is up to the interval that include the 99^{th} quantile of Y_i^c.

Table 3.5: aMSE of the conditional hazard models with the baseline hazard that follows Weibull with scale=3 and shape=0.50.

		aMSE of the baseline hazard				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.000792	0.003291	0.004309	0.005524	0.003567
200	0%	0.000546	0.001326	0.001886	0.002709	0.001846
500	0%	0.000398	0.000480	0.000745	0.001254	0.000867
1000	0%	0.000357	0.000239	0.000395	0.000677	0.000475
100	2.5%	0.000866	0.004288	0.005299	0.006047	0.004711
200	2.3%	0.000543	0.001441	0.002024	0.002673	0.001972
500	2.4%	0.000401	0.000525	0.000788	0.001049	0.000880
1000	2.4%	0.000355	0.000258	0.000409	0.000515	0.000470
100	5.9%	0.000844	0.004387	0.005309	0.005751	0.004584
200	5.7%	0.000546	0.001678	0.002197	0.002449	0.002423
500	5.7%	0.000413	0.000560	0.000869	0.000880	0.001170
1000	5.7%	0.000366	0.000281	0.000456	0.000433	0.000767
100	10.1%	0.000872	0.006254	0.007167	0.007027	0.008507
200	2.7%	0.000594	0.002602	0.003228	0.003287	0.004801
500	4%	0.000462	0.000708	0.001041	0.000322	0.002484
1000	10%	0.000425	0.000346	0.000546	0.000521	0.002016
100	25.5%	0.002001	0.009486	0.012651	0.011660	0.121134
200	25.5%	0.001335	0.005021	0.006488	0.006387	0.125883
500	25.6%	0.001014	0.001804	0.002365	0.002365	0.128161
1000	25.5%	0.000890	0.000988	0.001379	0.001379	0.125455

		aMSE of the conditional hazard at 8 combination of values for x^c and x^d				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.001549	0.005728	0.006985	0.011295	0.008368
200	0%	0.000917	0.002274	0.003457	0.005491	0.004436
500	0%	0.000627	0.000840	0.001581	0.002546	0.001865
1000	0%	0.000552	0.000417	0.001028	0.001562	0.001200
100	2.5%	0.001775	0.007231	0.007549	0.010635	0.009254
200	2.3%	0.000944	0.002501	0.003557	0.005026	0.004579
500	2.4%	0.000640	0.000914	0.001675	0.002191	0.001720
1000	2.4%	0.000550	0.000450	0.001109	0.001359	0.001077
100	5.9%	0.001620	0.007384	0.007583	0.009221	0.009813
200	5.7%	0.000950	0.002897	0.003607	0.004317	0.004213
500	5.7%	0.000681	0.000984	0.001832	0.001938	0.001638
1000	5.7%	0.000587	0.000492	0.001179	0.001193	0.001033
100	10.1%	0.001688	0.009874	0.009109	0.009583	0.012244
200	2.7%	0.001029	0.004385	0.004488	0.004652	0.005046
500	4%	0.000779	0.001253	0.002006	0.001980	0.002343
1000	10%	0.000702	0.000603	0.001360	0.001341	0.001741
100	25.5%	0.003391	0.014590	0.016039	0.015222	0.111126
200	25.5%	0.002196	0.007740	0.009154	0.008711	0.114156
500	25.6%	0.001709	0.003085	0.004748	0.004748	0.117009
1000	25.5%	0.001494	0.001693	0.003231	0.003231	0.116500

[1] The point of (X^c, X^d) that used to calculated the average aMSE are $(0,0), (1,0), (-1,0), (\overline{X}^c, 0), (0,1), (1,1), (-1,1), (\overline{X}^c, 1)$. The first point measures the hazard of the baseline group.

[2] The interval length is $b = 1$ and the number of intervals that used in the aMSE is up to the interval that include the 99^{th} quantile of Y_i^c.

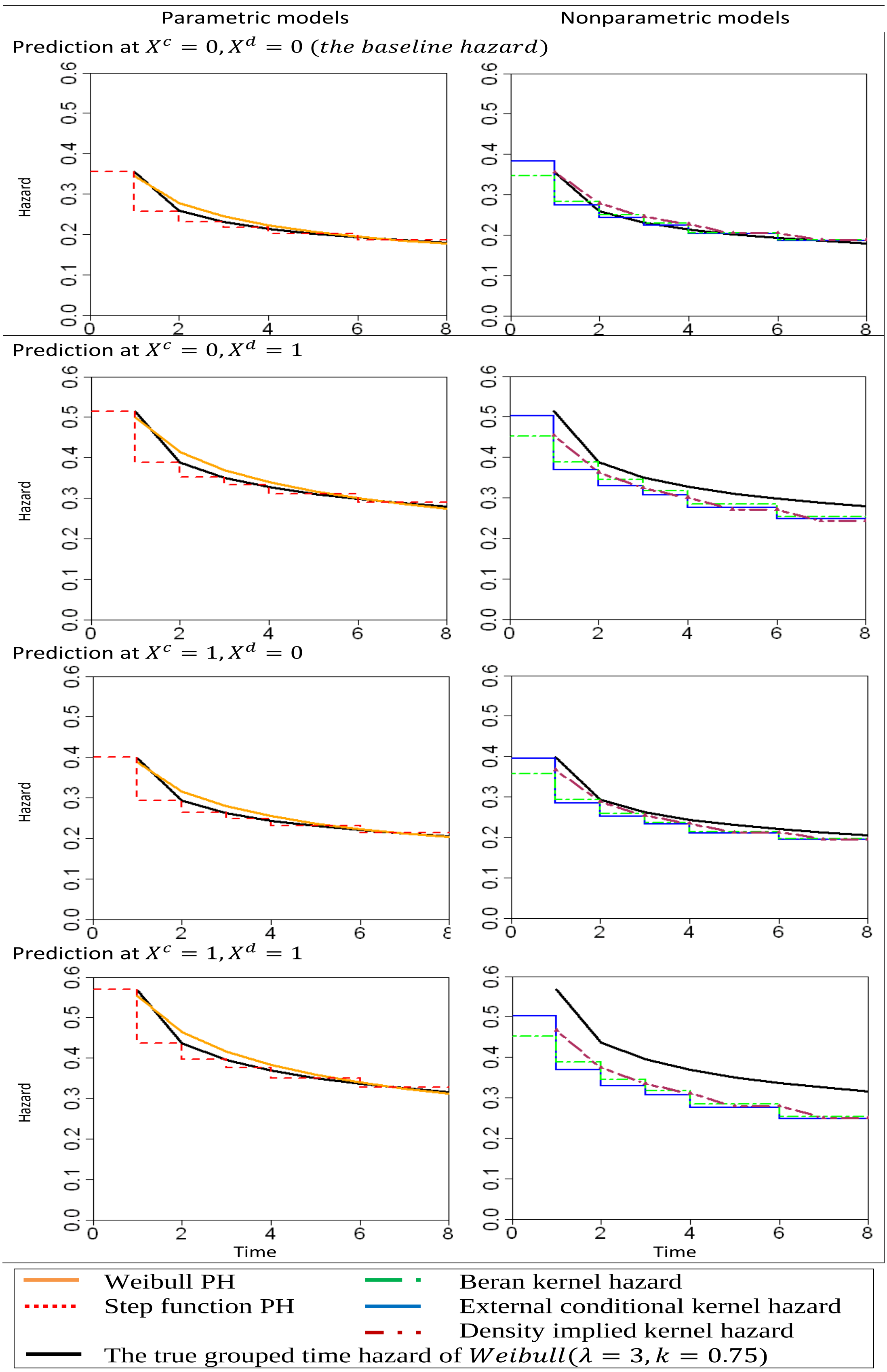


Figure 3.6: The predicted conditional hazard for the models with $Weibull(k = 0.75, \lambda = 3)$ baseline hazard and 0% censored observations.

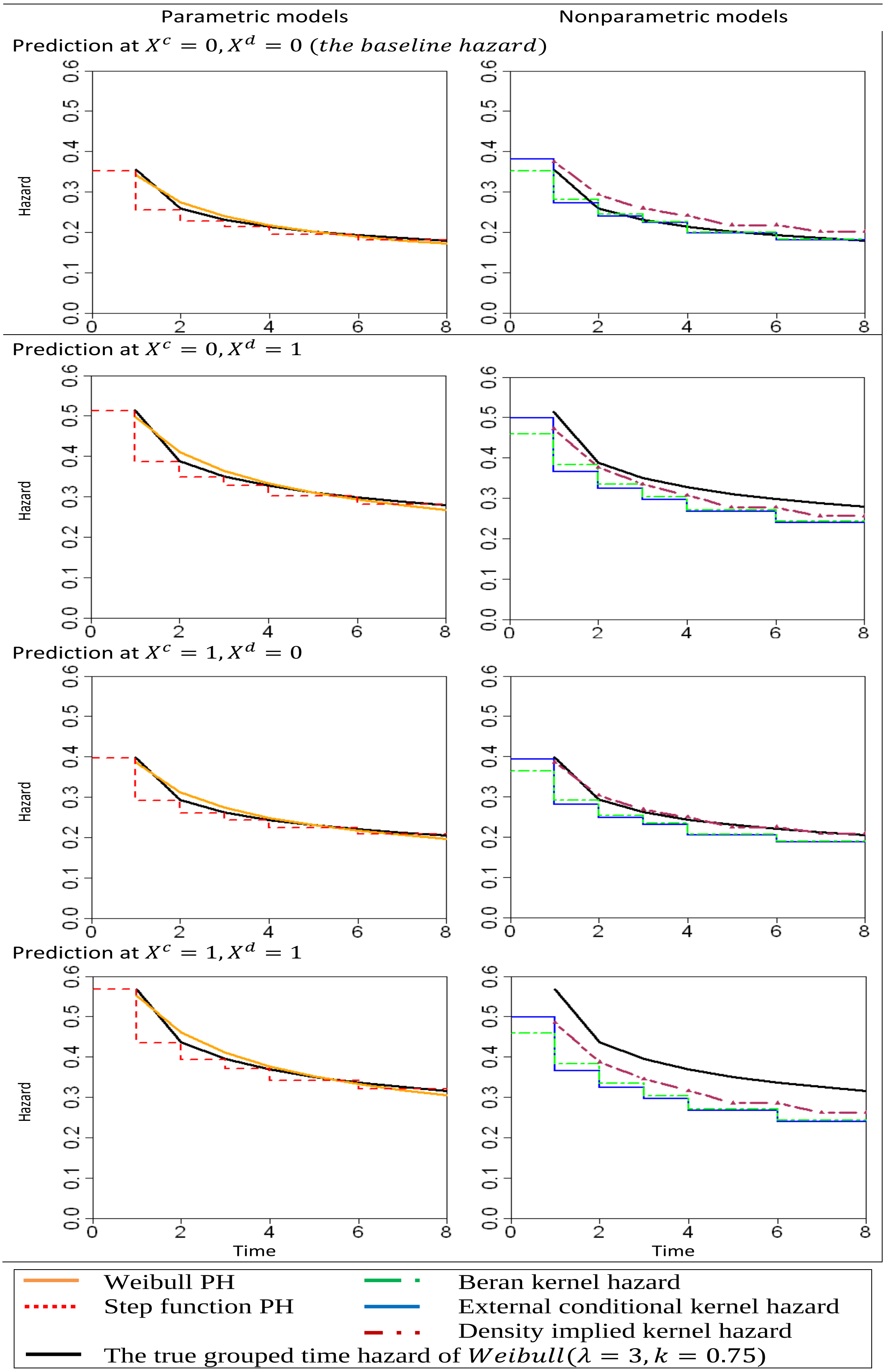


Figure 3.7: The predicted conditional hazard for the models with $Weibull(k = 0.75, \lambda = 3)$ baseline hazard and 5% censored observations.

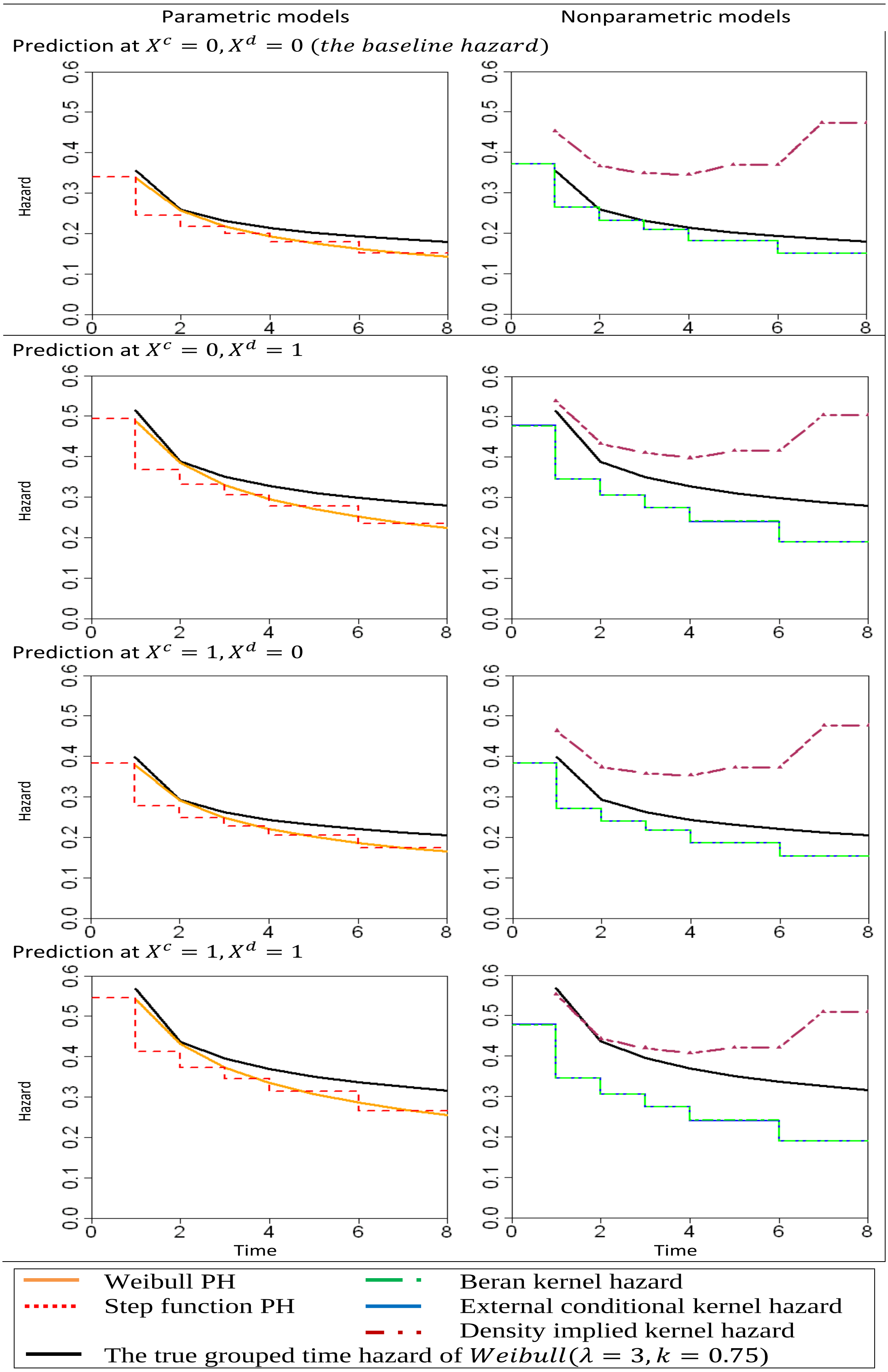


Figure 3.8: The predicted conditional hazard for the models with $Weibull(k = 0.75, \lambda = 3)$ baseline hazard and 25% censored observations.

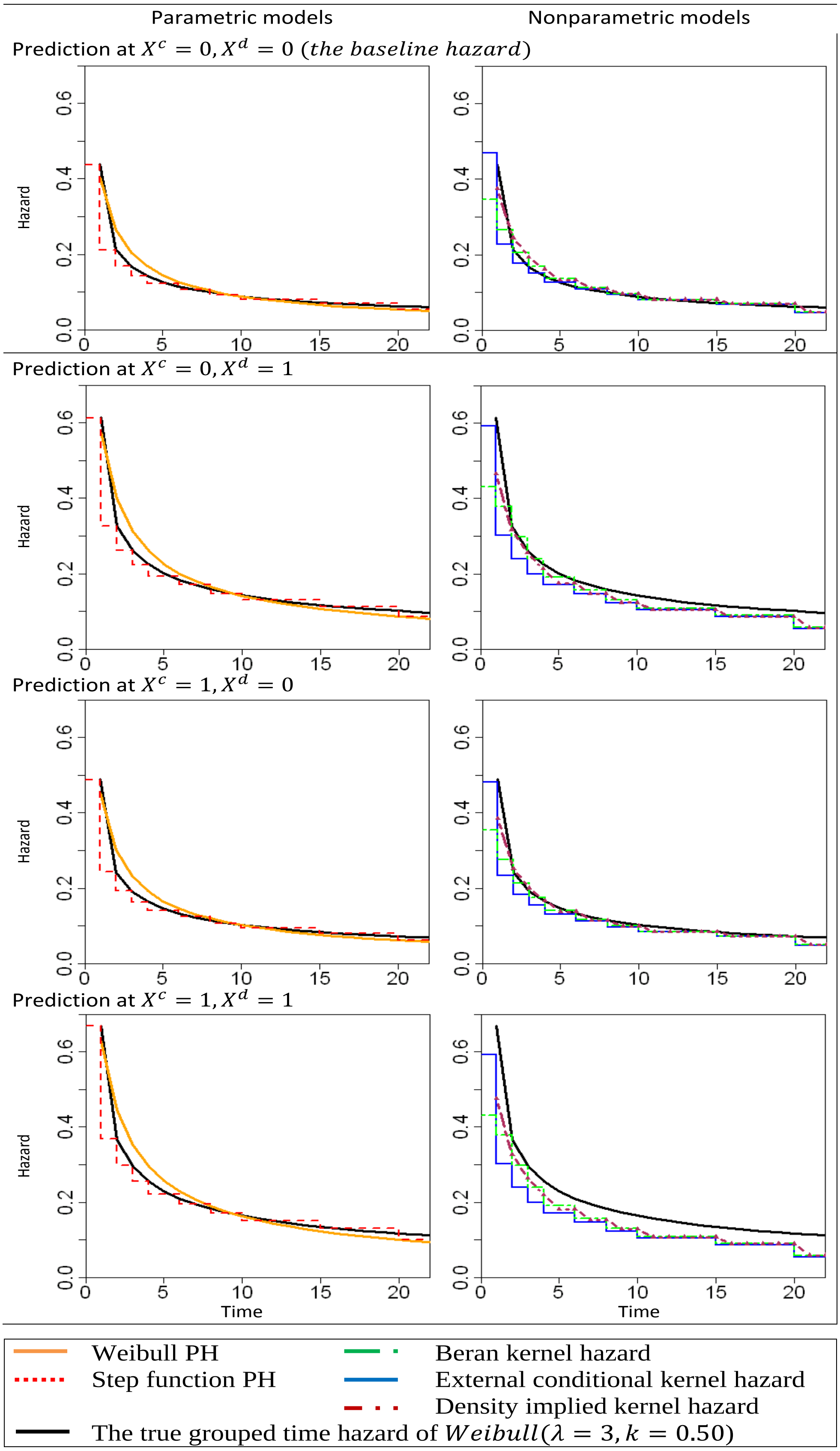


Figure 3.9: The predicted conditional hazard for the models with $Weibull(k = 0.50, \lambda = 3)$ baseline hazard and 0% censored observations.

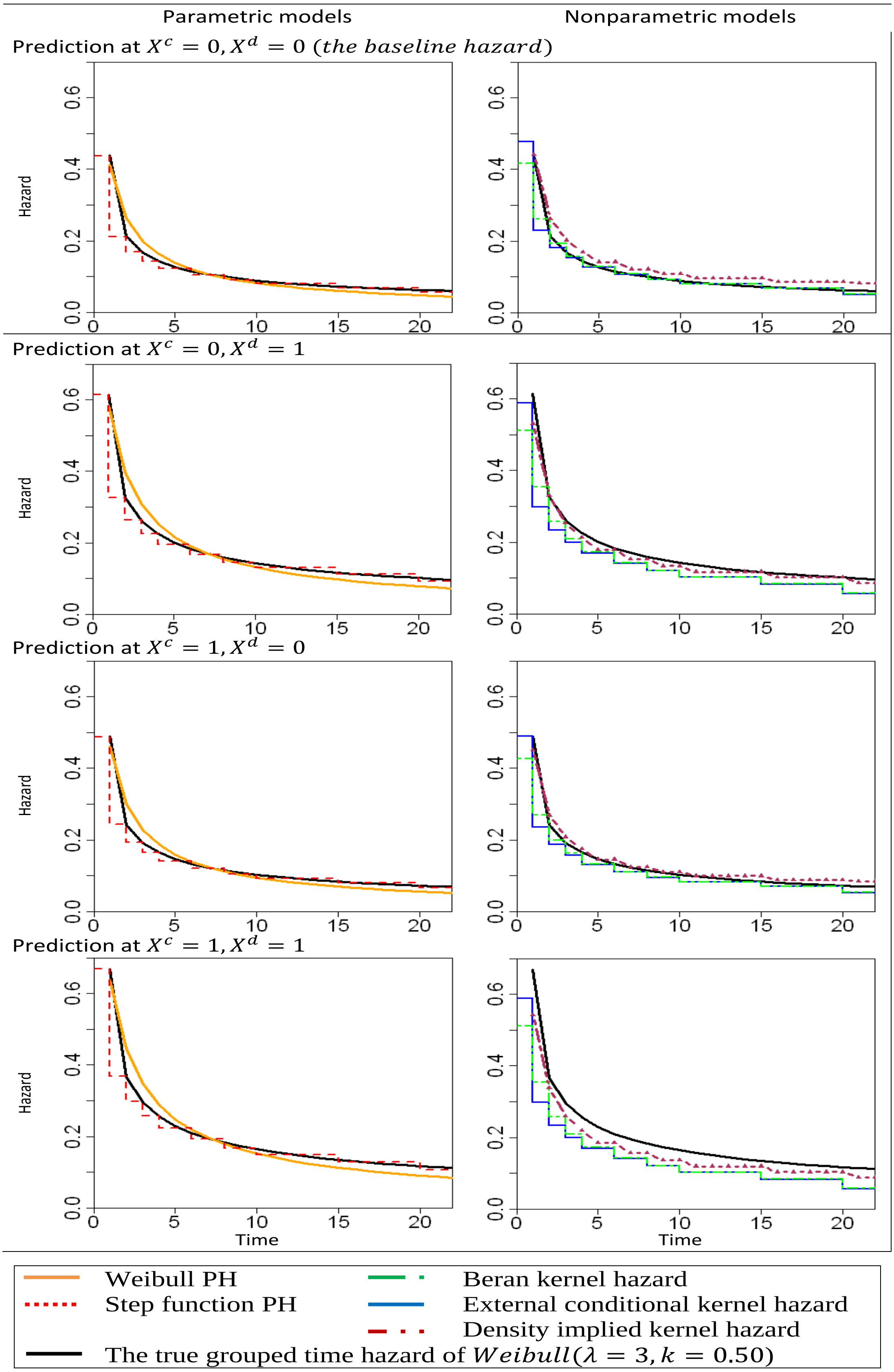


Figure 3.10: The predicted conditional hazard for the models with $Weibull(k = 0.50, \lambda = 3)$ baseline hazard and 5% censored observations.

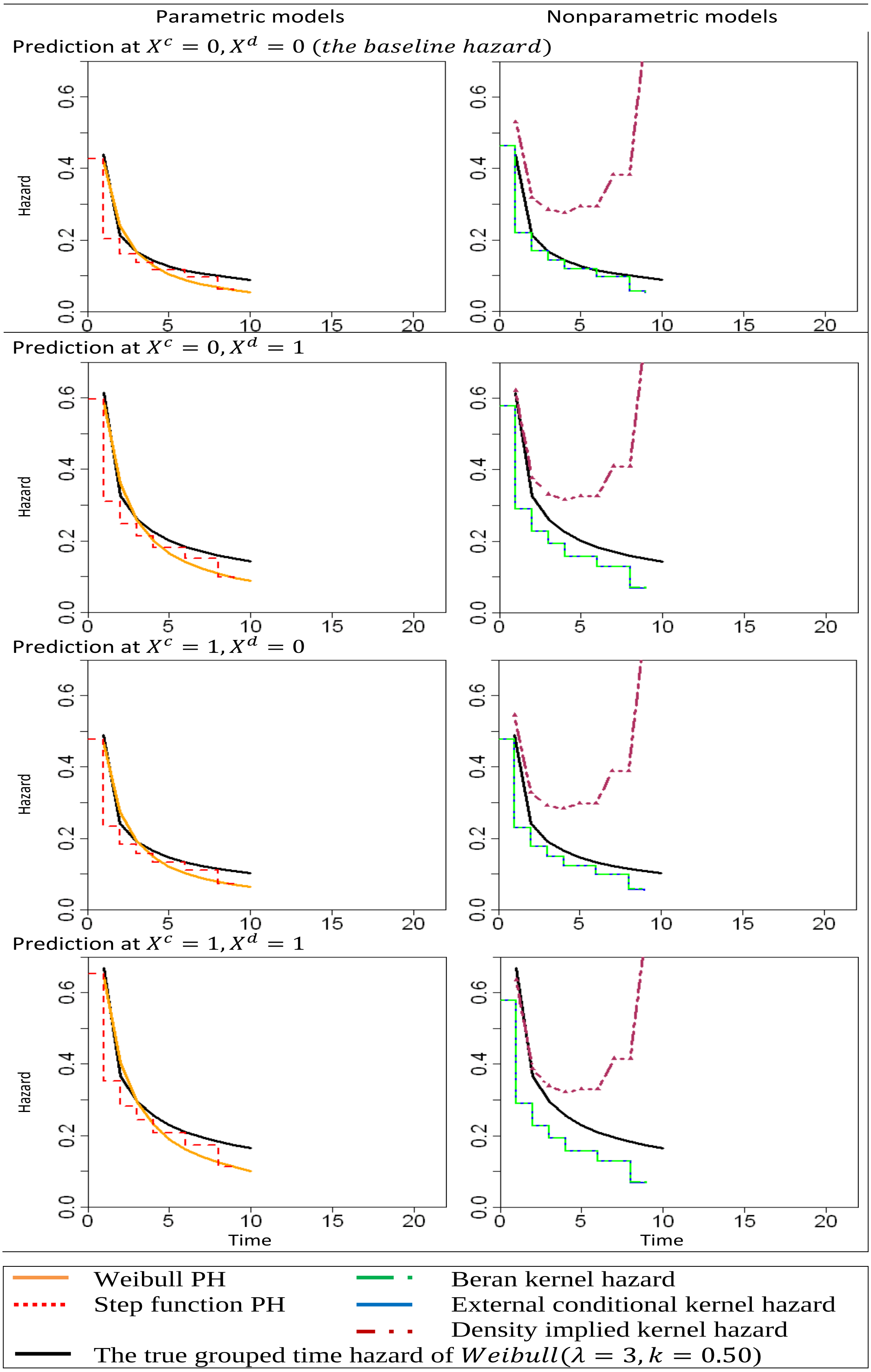


Figure 3.11: The predicted conditional hazard for the models with $Weibull(k = 0.50, \lambda = 3)$ baseline hazard and 25% censored observations.

Table 3.6: The bias for the conditional hazard for the models with the baseline hazard that follow Weibull with scale=3 and shape=0.75.

		The percentage bias of first episode				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0.0%	-3.4%	-0.3%	9.5%	-16.3%	-10.2%
200	0.0%	-2.1%	-0.1%	8.5%	-9.7%	-5.3%
500	0.0%	-2.7%	0.1%	6.8%	-3.5%	-1.0%
1000	0.0%	-2.9%	-0.2%	4.7%	-1.7%	-0.1%
100	2.2%	-4.0%	-0.8%	9.3%	-16.2%	-8.8%
200	2.1%	-3.7%	-0.7%	7.9%	-8.9%	-3.5%
500	2.2%	-3.5%	-0.7%	5.9%	-3.9%	0.1%
1000	2.1%	-2.9%	-0.2%	4.9%	-1.1%	1.9%
100	5.5%	-3.1%	-0.7%	9.3%	-13.4%	-4.5%
200	5.5%	-4.4%	-2.0%	7.0%	-8.0%	-0.6%
500	5.5%	-3.7%	-1.1%	5.4%	-2.3%	3.5%
1000	5.4%	-3.2%	-0.6%	4.6%	0.5%	5.5%
100	10.0%	-5.2%	-2.9%	7.4%	-11.3%	0.0%
200	9.9%	-4.4%	-2.1%	7.2%	-3.9%	5.8%
500	9.8%	-3.2%	-1.0%	6.1%	1.2%	9.6%
1000	9.4%	-3.6%	-1.4%	4.1%	2.0%	7.0%
100	25.9%	-6.0%	-4.9%	8.0%	1.8%	24.1%
200	25.6%	-5.6%	-4.7%	5.4%	3.5%	24.9%
500	25.7%	-4.9%	-4.1%	3.0%	2.7%	24.5%
1000	25.7%	-5.0%	-4.1%	1.0%	0.9%	22.9%

		Average of the absolute bias up to f the interval that include the 99^{th} quantile of Y_i^c				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0.0%	0.042394	0.073008	0.088750	0.083305	0.077528
200	0.0%	0.028042	0.046779	0.063335	0.060132	0.058278
500	0.0%	0.018230	0.030515	0.045094	0.043760	0.043269
1000	0.0%	0.014473	0.021388	0.034607	0.034056	0.033884
100	2.2%	0.041396	0.071953	0.089684	0.082626	0.078643
200	2.1%	0.028697	0.050382	0.064742	0.061469	0.059062
500	2.2%	0.018144	0.030031	0.044952	0.043742	0.042503
1000	2.1%	0.014249	0.021407	0.035459	0.034749	0.033680
100	5.5%	0.041932	0.073418	0.091269	0.083584	0.079667
200	5.5%	0.028177	0.049314	0.066620	0.062860	0.061060
500	5.5%	0.018437	0.030741	0.046158	0.044879	0.043789
1000	5.4%	0.014588	0.021812	0.036400	0.035742	0.035382
100	10.0%	0.042931	0.076560	0.092343	0.085193	0.085139
200	9.9%	0.029649	0.052257	0.067443	0.064391	0.064786
500	9.8%	0.019596	0.032366	0.048400	0.047883	0.049518
1000	9.4%	0.015530	0.022860	0.038210	0.037723	0.040350
100	25.9%	0.053652	0.098859	0.115988	0.111366	0.168271
200	25.6%	0.040491	0.066534	0.085390	0.084200	0.150649
500	25.7%	0.034574	0.044483	0.063765	0.063650	0.145367
1000	25.7%	0.031276	0.034839	0.051367	0.051337	0.145574

Table 3.7: The bias for the conditional hazard for the models with the baseline hazard that follow Weibull with scale=3 and shape=0.50.

		The percentage bias of first episode				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0.0%	-8.0%	-0.7%	7.3%	-39.1%	-29.9%
200	0.0%	-7.2%	-0.3%	7.3%	-31.5%	-23.2%
500	0.0%	-7.1%	-0.2%	6.1%	-21.6%	-15.2%
1000	0.0%	-7.2%	-0.4%	4.8%	-15.1%	-10.3%
100	2.5%	-6.9%	-0.3%	7.9%	-34.3%	-24.8%
200	2.3%	-6.7%	-0.2%	7.4%	-26.0%	-17.8%
500	2.4%	-6.3%	0.0%	6.6%	-15.6%	-9.4%
1000	2.4%	-6.5%	-0.2%	5.5%	-9.1%	-4.6%
100	5.9%	-6.6%	-1.2%	7.6%	-27.5%	-17.8%
200	5.7%	-6.2%	-0.7%	7.7%	-17.7%	-3.4%
500	5.7%	-5.7%	0.0%	7.4%	-6.3%	-0.8%
1000	5.7%	-6.2%	-0.5%	5.7%	-1.7%	2.3%
100	10.1%	-5.3%	-0.7%	9.4%	-18.0%	-7.9%
200	2.7%	-5.6%	-1.0%	8.3%	-8.8%	-0.9%
500	4.0%	-5.4%	-0.9%	6.8%	0.0%	5.5%
1000	10.0%	-5.3%	-0.7%	5.7%	2.9%	7.4%
100	25.5%	-4.3%	-2.7%	9.0%	2.9%	18.3%
200	25.5%	-4.7%	-2.9%	7.1%	6.1%	20.2%
500	25.6%	-4.8%	-3.0%	3.6%	3.5%	17.8%
1000	25.5%	-4.7%	-2.9%	1.6%	1.6%	16.2%

		Average of the absolute bias up to f the interval that include the 99^{th} quantile of Y_i^c				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0.0%	0.026296	0.054213	0.062051	0.066745	0.058949
200	0.0%	0.020644	0.035632	0.044853	0.048101	0.043417
500	0.0%	0.017471	0.022209	0.031246	0.034364	0.030970
1000	0.0%	0.016418	0.015704	0.024954	0.027542	0.025298
100	2.5%	0.028657	0.059212	0.063888	0.066672	0.061972
200	2.3%	0.021494	0.037089	0.045733	0.048083	0.044437
500	2.4%	0.018312	0.023053	0.032106	0.033745	0.030748
1000	2.4%	0.017287	0.016390	0.025986	0.027141	0.024804
100	5.9%	0.028342	0.060156	0.064236	0.064604	0.062860
200	5.7%	0.022665	0.031106	0.046816	0.047370	0.044642
500	5.7%	0.016176	0.024143	0.033558	0.033560	0.031259
1000	5.7%	0.019057	0.017176	0.026813	0.026643	0.025207
100	10.1%	0.030314	0.068068	0.069267	0.068339	0.072214
200	2.7%	0.024780	0.046196	0.050575	0.048828	0.052012
500	4.0%	0.022867	0.027179	0.035526	0.035094	0.038022
1000	10.0%	0.022473	0.019011	0.028890	0.028656	0.033075
100	25.5%	0.047010	0.090701	0.098648	0.095820	0.251130
200	25.5%	0.038742	0.064325	0.073451	0.072964	0.249570
500	25.6%	0.035393	0.041302	0.054177	0.054180	0.250981
1000	25.5%	0.033881	0.031035	0.044029	0.044029	0.252296

3.6 Conclusions

This research attempts to develop a kernel hazard estimator for discrete time single state transition model. There are two kernel estimators that are suggested in this research, an unconditional KH estimator and a conditional KH estimator. Both estimators are developed from the existing kernel estimators for single state continuous time transition model. The construction of the discrete time KH estimators is very similar to the construction of the continuous time hazard kernel estimators. The continuous time KH estimators use the information about the individuals in the risk set after each distinct transition time in the sample. The developed discrete time estimators, similarly, use the information about the individuals in the risk set just after each time index point in the underlying continuous time, the points that determine the grouped time intervals. The advantages of the discrete time KH estimators over the continuous time KH estimators are that they do not have source of Boundary Bias problem, are not affected by the interval censoring problem and are not affected by tie transitions.

The research also demonstrates that for the discrete time with equal length time intervals models, the Geometric Kernel function is appropriate to smooth the duration grouped time variable when the kernel estimators are used. The choice of the kernel function depends on the assumptions about the process. The Geometric Kernel is developed under the assumption that the number of intervals before the successful transition for individual i in the sample has a Geometric distribution. In the models with unequal length intervals, the assumptions regarding the transition in the process have to be considered to choose the kernel function of the duration grouped variable. The kernel estimators of the discrete time hazard are developed for the right censored data, however, they can be improved to the left truncated data as well. The asymptotic properties of the estimators are not discussed and left for future research. The discrete time kernel hazard estimators can be developed more to include method to control for the unobserved heterogeneity in the sample. The research suggests few kernel bandwidth estimation methods that can to be useful in this context.

The estimators that developed in this research are compared with the parametric proportional hazard discrete time estimator using Monte Carlo simulation method. The research creates a data generating process, DGP, that based on the grouping scheme of a continuous Weibull transition variable and on the assumptions of the PH framework. The kernel nonparametric hazard estimators show weaker performance than the parametric PH estimators. The reasons of the weak performance can be divided into two categories, reasons that are related to the DGP design in the research, and reasons that are related to the kernel method and the proportional hazard estimation framework.

The reasons that are related to the DGP and the simulation study in this research are:

1. The naïve bandwidths that are used in the simulation study are not optimal for

the kernel hazard rate estimators. The estimation of the optimum bandwidths is very tedious work in the Monte Carlo simulation study. The least-square cross validation risk optimum bandwidths are not founded because the variance and the bias formulas of the discrete time KH estimators are needed. This part is considered as a work that beyond the level of this research. The maximum likelihood cross validation risk (MLCV-Risk) optimum bandwidth can be estimated without further development, where the objective function of the MLCV-Risk hazard model is easier than the LSCV objective function. However, the computation of the MLCV-Risk bandwidths in the simulation study is very time consuming and is not feasible in our research time limits. The exclusion time of the simulation study to re-produce the simulations with the MLCV-Risk bandwidths is over 200 thousand hours. Which clearly justify that the use of the naïve bandwidths is the only feasible option in this case.

2. The largest sample size that used in the simulation study, $n = 1000$, is not large enough to demonstrate the performance of the kernel estimators adequately. Kernel estimators work better with large sample size, where the asymptotic properties of the estimators depend crucially on n. However, the hard computation cost that associated with the hazard estimators, which is generated from the estimation of the sub-survival functions in the model, makes the increase of the sample size to more than 1000 observations very difficult.

The reasons that related to the kernel estimation method and the proportional hazard estimation method are:

1. The kernel estimators are biased but consistent estimators, the asymptotic theory is necessary to show the properties of the estimators and how they converge to the true hazard functions. The discrete time KH estimators that are suggested in this research have asymptotic properties that are not discussed. The research draw the conclusions regarding the kernel hazard estimators based on the findings of the simulation study, which uses restrictive conditions and assumptions and excessive to the kernel estimators. The research focuses on the construction of the discrete time KH estimators, but due to the computation difficulties, done very little to allow those estimators be compared fairly with the proportional hazard parametric estimators.

2. The literature in conditional kernel hazard estimation is short and little informative.

3. The proportional hazard assumption is very restrictive, the assumption that the conditional hazard is a product of a baseline hazard and an observed heterogeneity function is not plausible empirically. However, the PH is the popular framework

to estimate duration models in econometrics, and we couldn't find an alternative estimation framework that applied to economic data. A link function conditional hazard estimators that relax the assumption regarding the baseline hazard is found in the literature of the parametric estimation of Survival Models in Younes and Lachin (1997). However, Younes and Lachin (1997) hazard estimator is for continuous time transition model, and is slightly discussed in the literature.

The kernel hazard estimators show a performance that indicates that they are promising class of estimators. However, the estimator that is suggested in this research has limitations, where the results are concluded from the simulation study only with little theoretical support, which shows the performance of the estimator under the DGP design only. Due to the hard computation of the KH estimator the naive bandwidths only are used, this limits use of the results in this research even more. The results on this chapter are only useful in that is highlights many of the advantages of estimating the hazard rate nonparametically, like the advantage that nonparametric estimation of the hazard does not need fully identifying the distribution of the underlying transition time and the discrete time KH estimator does not generate a Boundary Bias when the grouped time transition time is smoothed. The work in this research opens the door to do a lot of research in discrete time kernel hazard estimation.

This research is creative in that it develops a new discrete time kernel hazard KH estimators and design an original simulation study for grouped time duration models. The results and the findings of this research are important to economic duration estimation method and can be improved in many ways. The simulation study can easily be extended to include a simulation of the unobserved heterogeneity that follows Gaussian or Gamma distribution. The kernel hazard estimators in the simulation study can be compared with group regression AFT model in addition to the proportional hazard model. The work in this research can be extended to find the form of the optimum bandwidths of the discrete time KH estimators. The kernel hazard estimators, on the other hand, can be expanded also to comprise time varying independent variables and control for the unobserved heterogeneity in the duration model.

Chapter 4

Kernel Hazard Estimation of the Random Job Matching Model

Chapter Abstract

The research in this chapter applies the discrete time conditional kernel hazard estimator, that developed in the previous chapter empirically. The discrete time hazard is estimated for job-vacancies in Lancashire region in the UK, using a flow sample that covers the period from March 1988 to June 1992, and the duration is measured in weeks. The conditional external kernel hazard estimator is compared with the (mixed) proportional hazard estimators for discrete time single transition models. The labour-market random matching theory (RMT) is used to assess the performance of the estimation techniques. The research also introduces a maximum likelihood cross validation method to estimate the bandwidths of the conditional hazard estimators, and shows that it produces accurate results compared with naïve bandwidths.

The (mixed) proportional hazard estimators improves when the restrictive parametric assumptions regarding the baseline hazard form and the distribution of the unobserved heterogeneity are relaxed. But they show inconsistent results when the distribution assumption about the unobserved heterogeneity is changed, and produce results that contradict with the labour-market RMT. Kernel nonparametric estimators, on the other hand, show outstanding performance in term of the predicted hazard rate and the consistency of the results with the labour-market random matching theory. The kernel estimators show that the hazard rate of job-vacancies increase to about 3 times at the end of the flow sample period due to the increase of the number of job-seekers. The conditional external kernel hazard estimator capture the effect of labour market tightness in contrast to the (mixed) proportional hazard estimators.

Key Words: Nonparametric, random matching model, job-market tightness, job vacancies, hazard rate, survival function, grouped time hazard, Weibull hazard, proportional hazard, external kernel hazard, Beran kernel hazard, single state transition, , sub-survival function, naïve bandwidths.

JEL Classification: C14, C15, C24, C25, C51, C63, J10, J20, J41, J64.

4.1 Introduction

The objective of this chapter is to apply the discrete time kernel hazard estimation technique that is developed in Chapter 4. This chapter extends the methods of bandwidth estimation of the conditional kernel hazard estimators to the so-called the **maximum likelihood cross validation risk (MLCV-Risk) method**. The MLCV-Risk bandwidths are not used in Chapter 4 due to the high computation costs in their estimation. But in this chapter the computation cost is paid in an attempt to improve the kernel hazard estimators and enrich the discussion. The research uses data on weekly duration of job-vacancies drawn from the Lancashire Careers Service (LCS) administrative records. The proportional hazard (PH) estimators are compared with the conditional external kernel hazard (CEKH) and Beran kernel hazard (BKH) estimators in Chapter 3. The PH estimators are also improved by controlling for unobserved heterogeneity by using the mixed distributions proportional hazard estimation technique, known as the mixed proportional hazard (MPH) method.

The proportional hazard models are estimated with two specifications for the baseline hazard function. First, a fully parametric specification under the assumption that the underlying continuous time in the survival model follows Weibull distribution, and second, a semi-parametric specification of the grouped time baseline hazard, which is known as the step-function (piecewise) baseline hazard. The PH model without controlling for the unobserved heterogeneity is estimated for both specifications of the baseline hazard, but the MPH models are estimated with the step-function baseline hazard models only. The MPH models are estimated under two different assumptions: when the distribution of the unobserved heterogeneity is assumed to be a Gaussian distribution, and when it is assumed to follow a discrete distribution with a limited number of mass points. As shown empirically in Chapter 2 in the case of the conditional density estimation, parametric methods produce better results when restrictive parametric assumptions about the functional form in the model are relaxed. The MPH model with step-function baseline hazard and nonparametric discrete distribution of the unobserved heterogeneity is the parametric model with the most relaxed parametric assumptions in this research.

More specifically, the objective of this research is to examine the empirical results of the nonparametric conditional kernel hazard estimators relative to the results of the proportional hazard and the mixed proportional hazard parametric estimators, using labour-market data on duration of job-vacancies in weeks. The form of the hazard of job-vacancies is developed from the job-matching function of Pissarides (1990). The transformation of the matching function to a conditional hazard function conditioned on the number of job-seekers and the number of job-vacancies and other factors is illustrated in Section 5.2. In contrast to the comparison in the simulation study in the previous chapter, in this chapter the true hazard is not known and the sample is subject to an obvious unobserved heterogeneity problem. The theory regarding the job-matching function is informative about how the function responds to the changes in the number of job-seekers and the number of job-vacancies in the market pool, and shows that the rise in the number of job-seekers increases the probability of matching. .

The research deals with the unobserved heterogeneity problem in the data carefully and critically. In contrast to the parametric and semi-parametric estimation of duration models, where the unobserved heterogeneity problem is treated clearly by using mixture distributions, see Horowitz (1999), Baker and Melino (2000) and Abbring and Van Den Berg (2007), in kernel estimation treating UH ambiguous. The approaches to estimate kernel objects of mixture distribution is advanced theoretically and difficult to be handled empirically, in addition to that it is mostly covers continuous variables estimators. The our research focuses on the impact of the UH on the KH estimator and compares this with the impact of the UH in the parametric hazard estimator before and after controlling for the unobserved heterogeneity problem.

The chapter is organized as follows: Section 4.2 presents the theory regarding the labour-market matching function and the job-vacancies hazard rate development, then Section 4.3 we illustrate the econometric development of the parametric estimation technique and the kernel nonparametric estimation technique. The flow sample data and the sample statistics are discussed in Section 4.4. The models results and finally the conclusions are presented subsequently in Sections 4.5 and 4.6.

4.2 The Random Matching Model in the Labour Market

Random matching models aim to explain the relationship between job-seekers and employers in the trading sector of the job market. The trade sector is the first component of the labour-market before the production sector. The trade runs between the employers who offer job opportunities in the market and the unemployed workers who are interested in filling those jobs for wages. When a job is filled by an unemployed worker both the employer and the worker move to the production sector in the labour-market. However, the process corresponding the trade depends on macroeconomic and environmental factors as well as some individual characteristics regarding both unemployed workers and employers. This economic activity is a time consuming process constructed on labour supply and labour demand forces and leads to the determining of the wages in the labour-market. The matching function aims to specify the behaviour of employers and unemployed workers in the trading sector over time, and defines the matching between them as a time to event process that could be estimated with hazard modelling techniques. The employers that are looking for workers reveal this in the form of announced job-vacancies in the market. The unemployed workers, who are the job-seekers, respond to this by contacting with the employers according to the labour-market rules. There could be an intermediate body in the market like a job-office or an employment service that organizes the behaviour in the labour-market as a whole.

Most matching function focus on the numbers of job-vacancies and job-seekers to study this process in the labour-market.

The standard random matching model developed by Pissarides (1990) has a matching function that is constructed to be a well behaved function explains the transition from the trade sector to the production sector smoothly. The trade sector is constructed from a pool of job-vacancies that represent the demand side and job-seekers that represent the supply side. Stocks and flows in job-seekers and job-vacancies depend on the rate of transition from the trade to the production sectors in the market, and on the reverse transition. The flow is generated from loss of existing jobs in the production sector in addition to the entering of new workers and new job-vacancies to the market. The lose of existing jobs is assumed to have a steady effect in the model. The stock of job-vacancies and job-seekers on the other hand depends on the matching rate. A slow matching rate means a slow transition from the trade sector to the production sector and generates high stock of workers seeking jobs.

The matching function measures the average number of coordinated job-vacancies and job-seekers in the market per unit of time. Advanced labour-market matching model divide the job-seekers and the job-vacancies to stock and flow each, and estimate the matching function in implication of those components. The effect of the stock and flow in the matching function is examined in Andrews et al. (2011) and Forslund and Johansson (2007). The former paper shows that the flow of new vacancies is likely to increase the matching probability of the job-seekers in the stock by about three times. The model that use the total number of stock and flow job-vacancies and job-seekers is the basic standard random matching model of Pissarides (1990) and is the model that is considered in this research. The average matched number of job-vacancies per time unit, denoted by m, as shown by Pissarides (1990) is: $m = \mu(U, V)$, where U and V are the total number of unemployed workers and job-vacancies in the market respectively. μ is a well behaved function that has the same properties of the production function. The natural choice of $\mu(\cdot)$ is a Cobb-Douglas form function, $\mu = U^{\alpha}V^{\beta}$ with $0 < \alpha$, $\beta < 1$ and follows a Poisson process with a rate that equals the labour-market tightness $\rho \equiv V/U$, where the average number of contacts per vacancy is decreasing in ρ. The average number of contacts per vacancy is

$$m = \mu(U/V, 1) = U^{\alpha}V^{\beta-1}. \tag{4.1}$$

Under the random matching model the hazard rate for a vacancy to get filled is

$$\theta = a\mu(U/V, 1) = aU^{\alpha}V^{\beta-1}, \tag{4.2}$$

where a is the joint probability of offering a job and the probability of accepting the offer by a pair of contacted employer and job-seeker in the market. Then the log of the hazard is

$$\log\theta = \log(a) + \alpha\log(U) + (\beta-1)\log V + \varepsilon. \tag{4.3}$$

ε captures the effects of the unobserved factors in the model, which are assumed to be time invariant and independent of U and V.

For job-vacancy i the hazard of matching with a job-seeker in the market during the interval j^o is

$$\begin{aligned}
\theta\left(j^o \mid \mathbf{x}_{ijo}, \varepsilon\right) &= \Pr(y^c_{j^o-1} < T^c_i \leq y^c_{j^o} \left| T^c_i > y^c_{j^o-1}, \mathbf{x}_{ij^o}, \varepsilon\right), && (4.4)\\
&= \Pr(J^o_i = j^o \left| J^o_i \geq j^o \right., \mathbf{x}_{ij^o}, \varepsilon), && (4.5)\\
&= \frac{S(j^o - 1\left|\mathbf{x}_{ij^o}, \varepsilon\right) - S(j^o\left|\mathbf{x}_{ij^o}, \varepsilon\right)}{S(j^o - 1\left|\mathbf{x}_{ij^o}, \varepsilon\right)}, && (4.6)
\end{aligned}$$

where y^c is the underlying continuous transition time that measured in equal intervals $(0 = y^c_0, y^c_1]$, $(y^c_1, y^c_2]$, ...,, $\left(y^c_{B-1}, y^c_B\right]$. T^c_i is the i^{th} individual continuous duration time, which is not observed, instead the interval (the episode) of T^c_i is observed and denoted by J^o_i. The episode are defined from the continuous transition time as $J^o_i = 1$ if $y^c_0 < T^c_i \leq y^c_1$, $J^o_i = 2$ if $y^c_1 < T^c_i \leq y^c_2$,.. etc. Thus j^o is a discrete variable that denotes the episodes of transition and takes the values 1,2,..,B, it is illustrated more in the next section.

Using $\log\left[-\log\left(1-\theta\right)\right] \approx \log\theta$ and the baseline hazard in the interval j^o, $\tau_{j^o} = \log\left(\int\limits_{y_{j^o-1}}^{y_{j^o}} \phi_0(u)du\right)$, that is common for all job-vacancies, the hazard in Eq (4.3) can then be written as

$$\begin{aligned}
\log\left[-\log\left(1-\theta\right)\right] &\approx \log\theta, && (4.7)\\
&\approx \tau_{j^o} + \log(a) + \alpha\log(U) + (\beta-1)\log V + \varepsilon. && (4.8)
\end{aligned}$$

Because the approximation is very close, the equation is transformed to a complementary log-log link function form as follows

$$\theta = 1 - \exp\left[-\exp\left(\tau_{j^o} + \log(a) + \alpha\log(U) + (\beta-1)\log V + \varepsilon\right)\right]. \tag{4.9}$$

This equation is directly estimated by a discrete choice method with a cloglog link function, the coefficients α and β being interpreted as being the estimated elasticities of the labour supply and labour demand respectively.

The random matching model is directly estimated parametrically using the conditional hazard technique, whilst the mixed proportional hazard framework is used directly to estimate the function in Eq (4.9).

4.3 The Econometric Framework

The hazard rate is the **"instantaneous probability of leaving a state conditional on survival time"**, see Cameron and Trivedi (2005, page 576). The correlation between the conditional probability of leaving the state and the duration in the state is called the **duration dependence**. The matching of a job-vacancy with a job-seeker is a leaving of the state of being in the trade sector in the labour-market to the state of becoming active in the production sector. The duration dependence is negative if the probability of leaving the state decreases with the duration in the spell, and positive if the conditional probability increases with the duration in the spell. In rare cases, however, the dependence may be weak and undetectable by the model.

The data for labour-market random matching models have the regular problems of duration model data, so that the same sampling methods like the flow and stock sampling are used. The samples are mostly taken from administrative records of labour-market offices or government offices. The censoring problem occurs when the information about the duration in the state is incomplete for some individuals or vacancies in the sample. This causes a bias problem and uninformative in the model. The censoring complicates the estimation and reduces the effective sample size that used in the estimation. There are three types of censoring problems in survival analysis data: left censoring, interval and right censoring. The left censoring is when the beginning of the spell is not observed. It is the most difficult type to handle and causes serious bias in the duration model. It is advisable to eliminate the sources of the left censoring in the sample by choosing a sampling scheme that does not include the units that entered the state before the beginning of the observation period. The **flow sampling method**, for example, draws a sample from the unit entering the state after a specific time known as the **study start date**, so it does not include the sources of left censoring bias.

Interval censoring is defined as the case where the duration in the state is only known to continue between two points in time. In economic duration data the measurements are taken into time units like days, weeks, months, or in an unmeasured length intervals, which means that the data automatically correspond to the interval censoring problem. Although the process in many application runs originally in the continuous time, it may be difficult or even impossible to follow the units in the sample continuously to record or approximate the time of transition. Following the units through periods of time is one of the ways to overcome this problem. The transitions are then known to occur in an intervals only but the particular points in time at which they occurred are unknown. For this structure of the data the grouped time hazard model is also called interval censored data hazard model. In the grouped time hazard models the bias of the interval censoring is corrected by constructing an estimation technique using the functions of the continuous transition time at the boundaries of the intervals, which are also known as the time index points.

The last type of censoring, **the right censoring**, is when the transition is not observed. It causes a bias known as the right censoring bias, which is corrected in the models by including the distribution functions of the censored observations in the Maximum Likelihood function, see Klein and Moeschberger (2003) and Jenkins (2005).

Let f_{Y^c} and F_{Y^c} be the density and the distribution functions of the underlying continuous transition time Y^c. Let the right censoring continuous time, denoted by C^c, have a density r_{C^c} and a distribution function R_{C^c}. So for each observation i, the continuous time in the state is $T_i^c = min(Y_i^c, C_i^c)$ and has a distribution function $1-(1-F_{Y^c})(1-R_{C^c})$. The right censored observations are the observations that coincide with the value 1 in the indicator function $I\left(Y_i^c \leq C_i^c\right)$. The continuous time Y^c is observed in equal intervals each with length b. The intervals are numbered subsequently from 1 to B, where B is the last observed interval. This generates a discrete ordered random variable j^o that measures the duration in the state in intervals, j^o known as **episodes**. The intervals are defined theoretically with respect to the continuous duration time variable as $j^o = \left(y^c_{j^o-1}, y^c_{j^o}\right]$, where $y_1^c, y_2^c, \cdots, y_B^c$ are the time index points which are the upper bounds of the equal intervals. The index points are the key points that link the variables and the functions in the grouped and continuous forms. Note that after the last complete interval in the sample there is no upper bound where $y^c_{B+1} = \infty$ and the hazard in the episode $(y_B^c, \infty]$ equals 1 irrespective of the interval length or the number of intervals. In practice the sample end date is y_B^c and the observations that are still in the state up to this point are right censored.

Let $f\left(j^o\right) = f_{j^o}$ be the mass probability density of the discrete (discretized Y^c) transition time J^o. Then the distribution function of the continuous time Y^c and the discrete time J^o must be identical at the time index point, which are the ends of the intervals, as follows:

$$F_{J^o}\left(j^o\right) = \sum_{r=1}^{j^o} f_{J^o}\left(r\right) = \int_0^{y^c_{j^o}} f_{Y^c}(u)du = F_{Y^c}\left(y_{j^o}\right), \tag{4.10}$$

where $y^c_{j^o}$ are the time index points of the intervals. The hazard and the survival functions of the discrete (grouped) transition time, J^o, are defined in term of the survival function of the continuous time transition time, Y^c. The survival probability in the grouped time model in interval j^o is defined as the probability of surviving the interval and beyond, this means it is the complement of the distribution function at the end of the previous interval, i.e. $S\left(j^o\right) = 1 - F_{Y^c}\left(y^c_{j^o-1}\right)$. The discrete time hazard in each interval then follows directly from the survival function of the discrete time. The conditional survival function of the discrete time, J^o, is defined form the conditional distribution function of the continuous time Y^c, where $S\left(j^o|\mathbf{x}_{ij^o}, \varepsilon\right) = 1 - F_{Y^c}\left(y^c_{j^o-1}|\mathbf{x}_{ij^o}, \varepsilon\right)$, and $\mathbf{x}$ is the vector of the independent variables. This directly produces the hazard formula in Eq (4.6).

The conventional econometric method for discrete time hazard estimation is the **proportional hazard (PH)** framework of Jenkins (1995), which is flexible enough to control for the effect of the unobserved components. In contrast to the parametric methods of conditional hazard estimation, there are only few nonparametric kernel methods available. Thus to estimate a job-matching model, only the kernel hazard estimator that developed in Chapter 4 can be used. The following sub-sections examine the parametric method and the kernel nonparametric method for conditional hazard estimation in more details.

4.3.1 The Mixed Proportional Hazard Framework

The proportional hazard framework is the conventional technique to estimate the conditional hazard rate in grouped time parametrically. The PH framework assumes that the inter-individual differences in the sample have proportional impact on a common function that depends on time only known as the **baseline hazard**, $\theta_0(j^o)$. The baseline hazard function have the properties that it is:-

- A function on the elapsed time in the state only aims to estimate the duration dependence.
- The same for all the observations in the sample.
- Need not to be fully specified in the model.

Some proportional hazard estimation techniques apply the Partial Likelihood method or the Marginal Likelihood method to estimate the conditional hazard rate without defining a functional form to the baseline hazard, for example Cox regression which applies the Partial Likelihood. But those methods are used in continuous time conditional hazard estimation. For the discrete time case the baseline hazard can be nonparametrically defined in the model and the Maximum Likelihood method applies directly.

The inter-individual differences are divided into differences that are due to observed components in the sample and differences that are due to unobserved components. The observed components are the independent variables in the sample, $\mathbf{X}$, which have two types: time invariant independent variables and time varying independent variables. The contribution of the observed components to the baseline hazard is defined theoretically as a change in the scale of the transition time variable, but at the same time decomposable to a separate function known as the **observed heterogeneity function**, $\psi(\mathbf{X}) > 0$. This is estimated parametrically as a function on the independent variables only and needs to be fully specified in the model. The regular choice of the observed heterogeneity function in the literature is $\psi(\mathbf{X}) = \exp\left(\mathbf{X}^T\boldsymbol{\beta}\right)$. Clearly, the individuals in the sample with values of the independent variables that increase the hazard will have their conditional hazard rates estimate shifted upwards from the baseline hazard by a level determined by the values of independent variables and the estimated coefficients $\widehat{\boldsymbol{\beta}}$.

The proportional hazard functional form in the grouped time hazard case is the product of the baseline hazard and the observed heterogeneity functions, $\theta_0(j^o)\,\psi(\mathbf{X})$, which is transformed to the following form, see Jenkins (2005, Chapter 2):

$$\theta\left(j^{o}|\mathbf{X_i}\right) = 1-\exp\left\{-\exp\left[\log\left(\Lambda\left(y_{j^o}\right)-\Lambda\left(y_{j^o-1}\right)\right)+\mathbf{X}_i'\boldsymbol{\beta}\right]\right\}, \tag{4.11}$$

$$= 1-\exp\left\{-\exp\left[\log\left(\int\limits_{y_{j^o-1}}^{y_{j^o}}\phi_0(u)du\right)+\mathbf{X}_i'\boldsymbol{\beta}\right]\right\}, \tag{4.12}$$

$$= 1-\exp\left\{-\exp\left[\tau_{j^o}+\mathbf{X}_i'\boldsymbol{\beta}\right]\right\}, \tag{4.13}$$

where $\Lambda\left(y_{j^o}\right)$ is the cumulative hazard function of the underlying continuous time at the end of the interval j^o, $\int\limits_{y_{j^o-1}}^{y_{j^o}}\phi_0(u)du$ is the discrete time baseline hazard inside the interval. So, the discrete time baseline hazard $\theta_0\left(j^o\right)$ is defined from the continuous time baseline hazard $\phi_0\left(y^c\right)$.

The function in Eq (4.13) has a complementary log-log (cloglog) form which makes the estimation of the coefficients in τ_{j^o} part easy, but involves identification of the baseline hazard of j^o with respect to the parameters of the distribution of the continuous transition variable. However, τ_{j^o} may be fully identified when certain assumptions are enforced about $f\left(y^c\right)$, as in the case when it is assumed to be a Weibull density with shape parameter equals 1.

There are two methods to estimate the baseline hazard of the grouped time models. The first uses the logarithm of the grouped time j^o and the intercept term to define $\tau_{j^o}=\delta_0+\delta_1\log\left(j^o\right)$. This method identifies the continuous time baseline hazard through the values of the coefficients δ_0 and δ_1 if y^c has a Weibull distribution, particularly if the shape parameter is 1 (y^c is Exponential). However, the parameters of the Weibull density are not identified by the cloglog function if the shape parameter is different than 1, see the mathematical proofs in Appendix D. The second method uses a nonparametric step function to define the baseline hazard in the cloglog function. This method is known also as the piecewise baseline hazard, Jenkins (2005), and it is able to predict the baseline hazard of the grouped time better than the first method, but provides no information about the underlying continuous time density function. The step function is estimated by including time dummy variables in the cloglog function to approximate the baseline hazard in discrete time.

The mathematical proofs of Appendix D shows that for a discrete time hazard model with an underlying continuous time that follows Weibull with scale λ, shape parameter k and cumulative hazard function Λ_{Y^c} is:

$$\log\left[-\log\left(1-\theta\left(j^o\left|\mathbf{X}_i\right.\right)\right)\right] = \log\left[\left(\Lambda_{Y^c}\left(y^c_{j^o}\left|\mathbf{X}_i\right.\right)-\Lambda_{Y^c}\left(y^c_{j^o-1}\left|\mathbf{X}_i\right.\right)\right)\right], \tag{4.14}$$

$$= \log\left[\left(\frac{1}{\lambda\exp\left(-k^{-1}\mathbf{X}_i^T\boldsymbol{\beta}\right)}\right)^k\left(\left(j^o\right)^k-\left(j^o-1\right)^k\right)\right], \tag{4.15}$$

$$= \log\left(\frac{1}{\lambda}\right)^k+\log\left(\left(j^o\right)^k-\left(j^o-1\right)^k\right)+\mathbf{X}_i^T\boldsymbol{\beta}, \tag{4.16}$$

where the interval length is 1 and

$$\tau_{j^o} = \log\left(\frac{1}{\lambda}\right)^k + \log\left((j^o)^k - (j^o - 1)^k\right). \tag{4.17}$$

The log transformation aims to approximate the baseline hazard function:

$$\delta_0 + \delta_1 \log(j^o) \approx \log\left(\frac{1}{\lambda}\right)^k + \log\left((j^o)^k - (j^o - 1)^k\right) \tag{4.18}$$

clearly the first term is defined where $\delta_0 = \log\left(\frac{1}{\lambda}\right)^k$, the approximation is on the second terms $\delta_1 \log(j^o) \approx \log\left((j^o)^k - (j^o - 1)^k\right)$. In the Exponential case $k = 1$, then $\delta_0 = \log\left(\frac{1}{\lambda}\right)$ and $\log((j^o) - (j^o - 1)) = 0$. Then the estimated intercept coefficient is identified to the Exponential distribution scale parameter and the estimated slope is $\widehat{\delta}_1 = 0$, which is related to the shape parameter as $\widehat{\delta}_1 + 1 = 1$, which means a constant duration dependence. For the Weibull with shape parameter $k \neq 1$ the estimated slope coefficients is not identified in the model. But in the case of $k < 1$ the hazard is decreasing in time $\widehat{\delta}_1 < 0$ and $\widehat{\delta}_1 + 1 < 1$, when $k > 1$ the hazard is increasing in time $\widehat{\delta}_1 > 0$ and $\widehat{\delta}_1 + 1 > 1$. So a correct duration dependence in the grouped Weibull duration variable is estimated by the cloglog model.

Maximum Likelihood method is used to estimate the discrete time PH model. Jenkins (1995) shows that the likelihood function in Eq (4.13) is similar to the likelihood function of a panel data binary choice model with cloglog link function. The panel data binary model is estimated for a dependent binary variable d_{ij^o} that takes the value 1 only at the last episode for each individual only if the individual has a transition from the state and takes zero otherwise. The model is estimated by pooling the survival analysis dataset to an unbalanced panel data shape, and creates a new variable, a_{ir}, $r = 1, 2, \cdots, j_i^o$. For each individual in the sample a_{ir} measure the elapsed time in the state, it takes the values form 1 up to the number of the last episode that observed to the individual in the state, i.e $a_{i1} = 1$ and $a_{ij_i^o} = j_i^o$. The pooled sample discrete choice model is estimated with the covariates $log(a_{ir})$ and $\mathbf{X}_{ir}$ to estimate a cloglog with Weibull baseline hazard.

To estimate the discrete time model with step-function baseline hazard an episode specific dummy variables are used, a dummy variable that takes a value 1 for each individual at $a_{ir} = 1$ and zero otherwise, another dummy variable that takes values 1 when $a_{ir} = 2$ and zero otherwise, and subsequently up to the last dummy variable which equals 1 when $a_{ir} = max(J_i)$ and zero otherwise. The time dummy variables are then included in the cloglog link function with $\mathbf{X}_{ir}$ to estimate the model with step-function baseline hazard. The time dummies can be defined for grouped episodes instead to each episode as shown. For example, the episodes $10, .., 20$,which cover the underlying continuous transition time in the range $(y_{10}^c, y_{20}^c]$, can be grouped by define them by one dummy variable (step). The advantage of the grouped episodes step function over the step function of each episode is that it treats the diminishing number of individuals in the state in the late episodes in the sample better.

The log likelihood function of the discrete time proportional hazard model is, see Jenkins (2005, Chapter 6, page 72)

$$\log L = \sum_{i=1}^{n} \sum_{r=1}^{J_i^o} \left[d_{ir} \log \theta \left(r | \mathbf{X}_{ir} \right) + \left(1 - d_{ir} \right) \log \left(1 - \theta \left(r | \mathbf{X}_{ir} \right) \right) \right], \tag{4.19}$$

can be maximized in any conventional software that allows the estimation of discrete choice model with cloglog link function like STATA and R software.

The unobserved part of the inter-individual differences are known as the unobserved components in the proportional hazard model, v. The effect of v is defined as the **unobserved heterogeneity (UH)** in the duration models and they are assumed to be independent of time and not correlated with the independent variables, Van Keilegom and Veraverbeke (2001). The aggregate survival function is used to identify the PH model with the existence of the unobserved heterogeneity, as

$$\begin{aligned} S \left(y^c | \mathbf{x} \right) &= E_v \left[S \left(y^c | \mathbf{x}, v \right) \right], & (4.20) \\ &= \int \exp \left(-v \Lambda_0 \left(y^c \right) \psi \left(\mathbf{x} \right) \right) g \left(v \right) dv, & (4.21) \end{aligned}$$

where $g(v)$ is the unknown density function of the unobserved components. The model is estimated by making assumptions about $g(v)$ which have large impact on the likelihood function of the model. In the grouped time model the survival probability at the time index points of interval $j^o - 1$ is the survival probability for episode j^o and beyond, so

$$\begin{aligned} S \left(y_{j^o-1}^c | \mathbf{x} \right) &= \int \exp \left(-v \Lambda_0 \left(y_{j^o-1}^c \right) \psi \left(\mathbf{x} \right) \right) g \left(v \right) dv, & (4.22) \\ &= S \left(j^o | \mathbf{x} \right). & (4.23) \end{aligned}$$

This makes the identification of the distribution of the UH in the continuous time and the discrete time the same. For the identification of the hazard model the density function $g(v)$ in addition to the baseline hazard and the observed heterogeneity functions are assumed to be uniquely defined in the model. This characteristic is known as the **non-parametric identification** of the proportional hazard model, see Cameron and Trivedi (2005).

The hazard rate is defined as a product of the baseline hazard and the observed heterogeneity functions with v as

$$\theta \left(j^o | \mathbf{X} \right) = \theta_0 \left(j^o \right) \psi \left(\mathbf{X} \right) v, \tag{4.24}$$

with $v > 0$ and normalised so that $E[v] = 1$ by reparametrizing the density function $g(v)$.

The unobserved heterogeneity causes bias in hazard rate estimation. The effect of this bias in the observed heterogeneity and the duration dependence functions is studied by Andersen (1970), Lancaster (1979, 1990), Lancaster and Nickell (1980), Heckman and Singer (1984a,b,c), Ham and Rea Jr (1987), Meyer (1991), Van den Berg (2001) and Abbring and Van Den Berg (2007) among many. Ignoring the unobserved heterogeneity in the model causes serious bias toward zero in the coefficients of the independent variables, see Baker and Melino (2000), and causes an underestimation to the slope of the hazard rate, see Cameron and Trivedi (2005), where more negative duration dependence is estimated. The duration dependence that are affected by UH is known as **spurious duration dependence**. Due to the effect of UH, spurious duration dependence is more negative than the duration dependence when UH is controlled for in the model, because the individuals in the sample with higher values of v stay longer in the state. This relationship is known as the **weeding out** or the **sorting effect** problem, Heckman and Singer (1984b) and Van den Berg (2001).

The unobserved heterogeneity is defined in the model as individual specific random effects and time fixed components. The methods to control for UH use assumptions about the distribution of v added to the initial assumption about the distribution of the transition time variable y^c. The method of estimation is known as the **mixed proportional hazard (MPH) method**, each hazard model is defined by the combination of the distributions assumptions in $\theta_0(j^o)$ and $g(v)$ respectively, like Weibull-Gamma mixture or Exponential-Gaussian mixture. The density function of the unobserved components, $g(v)$, as mentioned by Cameron and Trivedi (2005), is arbitrary. When specified parametrically it should be assumed to be a function with a small number of parameters to avoid adding extra complications in the likelihood function. Heckman and Singer (1984b) and Heckman and Singer (1984c) developed a method to estimate the mixed model with nonparametric specification of the density of v, but adding more computations in the estimation technique. The mixing of the distributions directly affects the second moment of the hazard function by inflating the variance, Cameron and Trivedi (2005), this forces the researcher to trade between efficiency and bias in the mixed hazard model.

In the MPH literature the choice of the parametric form of $g(v)$ varies between the Gaussian density function and the Gamma density function, see Cameron and Trivedi (2005), Wooldridge (2002), and Lancaster and Nickell (1980). However, the asymptotic distribution of the UH, as shown by Abbring and Van Den Berg (2007), converges to a Gamma distribution for the individuals who remain in the state at the end of the study period (the survivals) in both continuous and discrete time MPH models, and in both univariate and multivariate duration models. Nicoletti and Rondinelli (2010) show that the mis-specification of the distribution of the unobserved heterogeneity causes bias in the estimated duration dependence function. But Heckman and Singer (1984b) argued that the parametric specification of the unobserved heterogeneity induces a lack of robustness in the model. Later, however, Heckman and Singer (1984c) introduce a **nonparametric maximum likelihood (NPML)** approach to estimate the mixed model which appears effective in estimating duration models. The NPML method assumes a discrete density function for v and approximates the distribution of the unobserved heterogeneity without assuming any parameters for $g(v)$ or any number for the mass points in advance. The NPML estimator of the hazard rate is widely used in estimating the duration models in economics. The mass points and their probabilities are estimated jointly with the coefficients of the observed heterogeneity function and the baseline hazard in the likelihood function via a hierarchical iterative procedure.

When the unobserved heterogeneity is controlled in the model the cloglog function in Eq (4.13) becomes

$$\theta\left(j^{o}|\mathbf{X}_i, v_i\right) \quad = \quad 1-\exp\left\{-\exp\left[\tau_{j^o}+\mathbf{X}_i'\boldsymbol{\beta}+\log\left(v_i\right)\right]\right\}. \tag{4.25}$$

Then the likelihood function of individual i takes the following form

$$L_i(\boldsymbol{\beta},\boldsymbol{\eta},\theta)=\int_{-\infty}^{\infty}\left[\prod_{r=1}^{J_i^o}\theta\left(J_i^o,\mathbf{X}_{ir}\left|v_i\right.\right)^{d_{ik}}\left[1-\theta\left(J_i^o,\mathbf{X}_i\left|v_i\right.\right)\right]^{1-d_{ik}}\right]g(v_i,\eta)dv_i, \tag{4.26}$$

where the vector $\boldsymbol{\eta}$ is the vector of the parameters of the unobserved heterogeneity density $g(\cdot)$. If the density $g(\cdot)$ is assumed as a density with many parameters this affects the length of the vector $\boldsymbol{\eta}$ and makes the optimization of the ML function more difficult. For the Heckman and Singer (1984c) method $g(\cdot)$ is assumed to be a discrete density and the integration in the ML function is replaced by a summation sign.

In practice the estimation of the discrete time MPH model is not straight forward and few software packages allow one to estimate any type of mixing. The Gaussian mixture is when a Normal density function is used for $g(\cdot)$. The estimation of the mixture model is then possible with random effects discrete choice estimation technique with the cloglog link function. This technique is available in STATA and R software directly, which makes the Gaussian mixture model the easiest type of MPH models to use practically. The alternative assumption, the Gamma distribution of v with Gamma density of the form

$$g(v;\boldsymbol{\eta})=\frac{\eta_1^{\eta_2}v^{\eta_2-1}\exp(-\eta_1 v)}{\Gamma(\eta_2)}, v>0,$$

makes the maximum likelihood function more difficult than maximums likelihood function of the Gaussian mixture model, but the MPH model then becomes consistent with the findings of Abbring and Van Den Berg (2007). Fewer packages allow the estimation of the Gamma mixture discrete hazard model; there is a package **PGMHZ8** of Jenkins

(1997)[1] in STATA, but there does not seem to be an equivalent package in R.

Finally, even fewer packages to estimate the MPH using Heckman and Singer (1984c) method, the Nonparametric Maximum Likelihood NPML method, are available. STATA is the only software that does this, with an option of two packages that available to use. The GLLAMM package of Rabe-Hesketh et al. (2001) in STATA is a powerful tool to estimate multilevel models and models of the panel data, see Rabe-Hesketh and Skrondal (2008). The second is the HSHAZ package of Jenkins (2006). The two packages produce identical results, but we prefer GLLAMM because it is faster than HSHAZ. The estimation of the MPH model using the Heckman and Singer (1984c) method in R software is available only in the **npmlreg** package using the EM method, which is not similar to the Newton Raphson Algorithm that used in the GLLAMM and HSHAZ packages.

To estimate the MPH model the function $g(\cdot)$ is assumed to be discrete function with limited number of mass points. The mass points are estimated using the Nonparametric Maximum Likelihood method of Heckman and Singer (1984c). The information criterion IC, the Akaike Information Criterion AIC and Bayesian Information Criterion BIC, are used to determine of the mass points in the estimated $g(\cdot)$ function, Rabe-Hesketh and Pickles (2002) and Rabe-Hesketh et al. (2004).

The hazard is estimated conditional on the independent variables: log total job-seekers, $\log U$, log number of job-vacancies, $\log V$, in addition to $year$ and $month$ dummies. Those dummy variables pick the seasonal variations in the hazard in the trading sector in the labour-market that are generated from the seasonal flow of job-seekers. The dummies of the sub-areas in the Lancashire region are included in the model also. The MPH is estimated with Gaussian mixture and with nonparametric mixtures. The model with Gamma mixture has been discarded due to the empirical difficulties relating to the slow convergence of the PGMHZ8 package. Also STATA is not available in the high performance computer cluster that is used for the long processing time and high levels of computation for the models in this research.

The term **covariate set** is used to describe the mixed type independent variables after they are reshaped by generating dummy variables to specify the discrete values of the categorical variables, as required to estimate the (M)PH models, see the discussion about the differences in inputting discrete categorical variable between parametric and nonparametric methods in Chapter 2. The reshaped covariate set contain categorical variables only, where the $year$, $month$ and lad variables are divided into dummy variables. The base group for the $year$ and $month$ variables are set to the earliest year in the sample (1988) and January month respectively. This generates 15 dummy variables aiming to pick up the seasonal variation in the matching process in the labour-market. The basegroup of the lad sub-areas is East Lancashire. The lad sub-areas are three, so two dummy variables are used. Hence the mixed independent

[1]The package PGMHZ8 compatible with STATA 8 and latter versions, the article Jenkins (1997) refers to the older version.

variables $[\log U, \log V, year, month, lad]$, which are a vector of length 5 for each observation, are reshaped in the (M)PH models to a covariate set $[\ \log U, \log V,\ d_{1989}, \cdots,$ $d_{1992}, d_{Feb},\ \cdots, d_{Dec},\ d_{Central}, d_{East}\ ]$ with length 19 for each observation. The estimated coefficients of the covariates give the predicted observed heterogeneity function in the model.

4.3.2 The Kernel Hazard Approach

This research uses the conditional kernel hazard (KH) estimators as developed in Chapter 3, the conditional external kernel hazard (CEKH) and Beran kernel hazard (BKH). The discrete time KH estimator are extended formulas from the continuous time KH of Watson and Leadbetter (1964a,b) and Beran (1981), respectively, to discrete grouped time and right censored observations as illustrated in Chapter 4. The CEKH estimator at episode j^o conditional on the vector of mixed type independent variables $\mathbf{x}$ takes the form

$$\widehat{\theta}^{CEKH}\left(j^o\left|\mathbf{x}\right.\right) = \sum_{i=1}^{n} l\left(J_i^o, j^o, \widehat{\gamma}_0\right) \frac{\mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}})\ \triangle_i}{1 - \sum\limits_{i:J_i^o<j^o} \mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}})}, \tag{4.27}$$

where $l\left(\cdot\,,\cdot\,,\cdot\right)$ is an ordered discrete variable kernel function used to smooth the duration time in the state. It is defined as a Geometric Kernel function due to the advantages that shown in Section 4.2 in Chapter 3. The Geometric Kernel takes the following form:

$$l\left(J_i^o, j^o, \gamma\right) = \begin{cases} 1-\gamma & \text{if } \left|J_i^o - j^o\right| = 0 \\ \frac{1}{2}\left(1-\gamma\right)\gamma^{\left|J_i^o - j^o\right|} & \text{if } \left|J_i^o - j^o\right| \neq 0 \end{cases}, \tag{4.28}$$

Beran kernel hazard, on the other hand, takes the following form in grouped time models:

$$\widehat{\theta}^{BKH}\left(j^o\left|\mathbf{x}\right.\right) \quad = \quad \frac{\sum\limits_{i:J_i^o=j^o} \mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}})\ \triangle_i}{1 - \sum\limits_{i:J_i^o<j^o} \mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}})}, \tag{4.29}$$

where the duration time variable is not smoothed in the model.

$\mathbf{A}(\mathbf{x}, \mathbf{X}_i, \widehat{\mathbf{h}})$ are the Nadaraya-Watson weights. The term in the denominator of each estimator is the estimated sub-survival function for the vacancies in the trading state of the labour-market (the units at risk) at the beginning of interval j^o. In the numerators the Nadaraya-Watson weights are for the uncensored vacancies in interval j^o, where $\triangle_i$ is the right uncensoring dummy variable that equals 1 for transitions and 0 for right censored observations. The smoothing technique embodied in the Nadaraya-Watson weights is the mixed type variables smoothing technique that presented in Li and Racine (2007). The weights are generated from the ratio of the kernel products of the independent variables in the model,

$$\mathbf{A}(\mathbf{x}, \mathbf{X}_i, \mathbf{h}) = \frac{n^{-1}W(\mathbf{x}, \mathbf{X}_i, \mathbf{h})}{\sum_{i=1}^{n} n^{-1}W(\mathbf{x}, \mathbf{X}_i, \mathbf{h})}. \tag{4.30}$$

The general form of the product kernel functions for a model that includes q continuous variables, each smoothed by a continuous kernel function $w\left(\frac{X_{is}^c - x_s^c}{h_s}\right)$, and p discrete variables, each smoothed by a discrete kernel function $l\left(X_{ir}^d, x_r^d, \gamma_r\right)$, is

$$W(\mathbf{x}, \mathbf{X}_i, \mathbf{h}) = \prod_{s=1}^{q} h_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{h_s}\right) \prod_{r=1}^{p} l\left(X_{ir}^d, x_r^d, \gamma_r\right). \tag{4.31}$$

The discrete variables, X^d, can either be unordered or ordered categorical variables or a mixture of both. The scalar $\widehat{\gamma}_0$ in the CEKH estimator in Eq (4.27) is the cross-validation bandwidth of the discrete time variable. The vector of the independent variables cross-validation bandwidths, $\widehat{\mathbf{h}}$, has length $q + p$ and contains a sub-vector of the continuous independent variables bandwidths $\widehat{\mathbf{h}}_{\mathbf{x}^c}$ and a sub-vector of the discrete independent variables bandwidths $\widehat{\boldsymbol{\gamma}}_{\mathbf{x}^d}$, so $\widehat{\mathbf{h}} = \left[\widehat{\mathbf{h}}_{\mathbf{x}^c}, \widehat{\boldsymbol{\gamma}}_{\mathbf{x}^d}\right]$.

The mixed type independent variables in the matching model, $[\log U$, $\log V$, $year$, $month$, $lad]$ have also been reshaped for the KH estimators, but by merging the $year$ and $month$ variables into a new generated variable $dmonth$ that gives the dates in months in the sample as distinctive ordered values. Each combination of values in the $year$ and $month$ variables gives a date in the sample in months. In the parametric reshaped data each month in the sample period is equivalent to a unique vector of combination of 0-1 values generated by the 15 dummy variables of $year$ and $month$. Then the dates in months from the beginning of the flow sample period up to the end can be represented as discrete, increasing ordered values, analogous to their representation as a unique combination of 0-1 values in the dummy variables. The new variable is a discrete ordered variable that counts the length of time in the flow sample in months since the start date of the observation period, and so has a similar construction to the duration variable in the state. The duration time in the state is interpreted as the number of weeks that a vacancy stays in the trading state in the labour-market, whiles the $dmonth$ variable is interpreted as the number of months that passed since the start of the observation in the flow sample. Both variables are then smoothed by the Geometric kernel function.

However, the $dmonth$ variable is related to calendar dates, and it is only interpretable when it is converted back to the calendar month of the year. As in $year$ and $month$ dummy variables in the parametric MPH model, the discrete ordered variable $dmonth$ aims to capture the seasonal variations in the model. In contrast, the duration time in the state in weeks is independent of calendar dates, when the sample is taken from the vacancies that are announced at the dates between the start and end of the sample period. The reshaping to decrease the number of variables in the kernel estimators is an effective technique, where it reduces the number of the independent variables in the kernel estimator. This reduces the chance of getting into the curse of dimensionality problem, and shortens the computation time of the model without losing any information from the sample. In the parametric method however it has the reverse effect. The newly generated variable has 52 distinct ordered values, and need 51 dummy variables to be included in the MPH model, which would make the MPH difficult to interpret.

The number of continuous independent variables in the kernel hazard estimators is $q = 2$, $\log U$ and $\log V$, and the number of the discrete variables is $p = 2$, $dmonth$ and lad among which only lad is unordered categorical variable. The Second-Order Gaussian Kernel function,

$$w\left(z\right) = \left(2\pi\right)^{-\frac{1}{2}} \exp(-z^2/2), \tag{4.32}$$

is used to smooth the continuous independent variables, and the Aitchison and Aitken (1979) unordered discrete variable kernel function

$$l\left(X_i^d, x^d, \gamma\right) = \left[1-\gamma\right]^{d_i} \left[\frac{\gamma}{(c-1)}\right]^{1-d_i}, \quad d_i = 1\left(X_i^d = x^d\right), \tag{4.33}$$

is used to smooth the lad variable. The Geometric kernel function is used to smooth $dmonth$ variables as mentioned earlier.

Bandwidth Estimation Method

As in all kernel estimators, the method to estimate the bandwidths in the model is more crucial to the quality of the results than the choice of the kernel functions of the variables in the model. The functional form of the optimum bandwidths of the kernel hazard estimator is not provided in the discussion in Chapter 4. A list of the possible methods to estimate the bandwidths is presented in that chapter with an argument which states that the MLCV method is valid to use to estimate the bandwidths for the discrete time kernel hazard estimator. In this chapter we apply this argument to improve the quality of the kernel hazard estimators.

The application of the MLCV method to the hazard rate is suggested by Tanner and Wong (1984). The method is described as an application of the MLCV to the hazard function. The produced bandwidths are better than the naïve bandwidths because they are estimated from the hazard function directly and consider the sub-survival functions in the KH formulas. But the MLCV method has a traditional problem with kernel estimators, where it tends to oversmooth estimators of the functions that are increasing or decreasing, or functions of the variables that have high skewness, like the sub-survival function in the hazard rate. In our empirical resutls the naïve bandwidths are also included, but they are used as starting values in the objective function of the MLCV method. The value of the likelihood function indicates whether the estimators are improved when the MLCV method is used or not.

The naïve bandwidths are just the MLCV bandwidths of the conditional density of the duration time conditioning on the mixed independent variables. The objective function of the conditional density MLCV bandwidths is shown in Hall et al. (2004) and Li and Racine (2007), with the leave-one-out method, as follows

$$\begin{aligned} L &= \log \sum_{i=1}^{n} \widehat{f}_{-i}\left(j_i^o | \mathbf{X}_i\right), \\ &= \sum_{i=1}^{n} \log \left(\frac{\sum\limits_{i'=1; i' \neq i}^{n} (n-1)^{-1} l\left(J_i^o, J_{i'}^o, \widehat{\gamma}_0\right) \prod\limits_{s=1}^{q} h_s^{-1} w\left(\frac{X_{is}^c - X_{i's}^c}{h_s}\right) \prod\limits_{r=1}^{p} l\left(X_{ir}^d, X_{i'r}^d, \gamma_r\right)}{\sum\limits_{i'=1; i' \neq i}^{n} (n-1)^{-1} \prod\limits_{s=1}^{q} h_s^{-1} w\left(\frac{X_{is}^c - X_{i's}^c}{h_s}\right) \prod\limits_{r=1}^{p} l\left(X_{ir}^d, X_{i'r}^d, \gamma_r\right)} \right). \end{aligned} \quad (4.34)$$

The function $\widehat{f}_{-i}\left(j_i^o | \mathbf{X}_i\right)$ is the conditional density of the discrete duration time at J_i^o on $\mathbf{X}_i$, using all the observations in the sample except the i^{th} observations. This means that the distance between all the observations of the duration time from J_i^o, and the distance of each independent variable in $\mathbf{X}$ from $\mathbf{X_i}$, which are the columns of the matrix $\mathbf{X} - \mathbf{X}_i$, are locally weighted with kernel functions that are specified for each variable after dropping the distances corresponding to the i^{th} observation. The number of the locally weighted observations is $n-1$, as expected with the leave-one-out method. Note that the leave-one-out method is developed from the jackknife bootstrap method, see Chapter 2 for more information.

The so-called MLCV-Risk bandwidths are used to refer to the extension of the MLCV method for the continuous time hazard of Tanner and Wong (1984) to the discrete time KH estimators. The maximum likelihood function of the discrete time hazard function in Jenkins (2005) is used. The maximum likelihood function of the discrete time hazard is presented in Jenkins (1995) and shown in discrete choice binary model maximum likelihood form to allow to use the cloglog link function to estimate the hazard parametrically. We start from the same original form of the maximum likelihood function as presented in Jenkins (2005, Chapter 6, Eq (6.10)), except that we will applies the leave-one-out method in the conditional kernel hazard in the formula. We refer to the method as the MLCV-Risk method to distinguish it from the MLCV of the kernel conditional density estimator that is used to estimate the naïve bandwidths. The objective function of the MLCV-Risk method is as follows:

$$ML\left(\gamma_0, \mathbf{h}\right) = \sum_{i=1}^{n} \left[\triangle_i \log \left(\frac{\widehat{\theta}_{-i}^{(\cdot)}\left(J_i^o \left| \mathbf{X}_i\right.\right)}{1 - \widehat{\theta}_{-i}^{(\cdot)}\left(J_i^o \left| \mathbf{X}_i\right.\right)} \right) + \sum_{r=1}^{J_i^o} \log \left(1 - \widehat{\theta}_{-i}^{(\cdot)}\left(r \left| \mathbf{X}_i\right.\right) \right) \right], \quad (4.35)$$

where $(\cdot)$ denotes the CEKH or BCH estimator.

The difference between the MPH method and the KH method is not only in the reshaping of the independent variables, to generate dummy variables for the parametric estimators or merge categorical variables to apply the nonparametric kernel estimation method. But also on the number of the observations that are used in each method. The MPH method uses the pooled sample data-shape to estimate the hazard, whilst the kernel hazard method uses the survival analysis data-shape. The survival analysis data-shape is the data shape of continuous time kernel hazard estimators, so that the discrete time kernel hazard estimators use this data-shape. The survival analysis data-shape compared with the pooled sample (unbalanced panel) data-shape is just a realisation of each unit in the sample at the last episode that observed for it in the state. However this is only a rough description to the difference. The information about the duration in the state is not changed by this reshaping, but the information about the values of the time varying independent variables before the last episode is lost.

Consequently, the current kernel hazard estimators are not able to handle time varying independent variables. Conditional kernel hazard estimators with time varying regressors are up till now discussed very briefly in the literature, as in Nielsen and Linton (1995), but all the sources that used to develop CEKH and BKH are silent about time varying regressors. One can argue that the kernel hazard estimators in this chapter and in Chapter 3 can be extended to smooth time varying independent variables with only an appropriate modification of the local weights in the sub-survival functions, say for example from the regular Nadaraya-Watson weights to possibly more sophisticated type of weights. The literature of kernel methods in time series provides families of local weighting functions that are attractive to try here. This issue is one of the future research focuses of the discrete time kernel hazard estimators.

4.4 The Data

4.4.1 Lancashire Careers Service Dataset

A flow sample from the job vacancies in the administrative records of Lancashire Careers Service (LCS) is drawn to estimate the random matching model in this research. The sample covers the period from March 1988 to June 1992 in three sub-areas and follows the job-vacancies and the job-seekers weekly. The LCS was a Government-funded network which acted as an employment service for youth in the age group 15-18. Individuals in this age group have few options for their future careers, they can either continue in school (be in education), move to a government training program, or leave school to work. In the last option they spend time in the trade sector of the labour-market searching until they find a job and start work. The LCS therefore worked as employment service, or a job center, for youth in the area who are seeking job as well as for employers who are looking for workers/employees. In contrast to the adult job service, the notification of job vacancies for youth was compulsory for all employers in the region, where the Government aimed to monitor the working experience of the youth population in the UK through this service, see Andrews, Bradley and Stott (2001) for more details.

The LCS is a rich database and provides sufficient information to experiment with many types of labour market models including the random matching model. The sample of our research is used in Andrews, Bradley and Upward (2001), Andrews, Bradley and Stott (2001) and Andrews et al. (2011) among many other papers. The job vacancies that announced in LCS for youth job-seekers account for about 30% of all job vacancies in the region during the period of the sample. Each vacancy is followed in the sample until a matching with an unemployed young job-seeker is observed, or, becomes right censored. A youth in employment, re-unemployment, government training and education is dropped from the sample, so only unemployed youth are counted on the number of job-seekers variable. The vacancies are right censored if they were withdrawn from the market or stayed open until the end of the observation period. Hence, the vacancies that are filled by unemployed young job-seekers only are included in the flow sample in addition to the censored vacancies. The duration in the state is measured in weeks starting from the week that the vacancy was announced in the LCS office until the matched job-seeker actually started working in the position.

The motivation of our research, in addition to comparing the results of kernel hazard estimator with the MPH estimator, is to study the effect of the job-market tightness on the duration of the job-vacancies and the hazard rate for making transition from the trading sector in the job-market and check the argument that kernel estimators are not by the unobserved heterogeneity problem as much as the parametric estimators. Kernel hazard estimators is developed for homogeneous models, i.e when no unobserved heterogeneity exists in the transition processes. Thus our research focuses on examining the impact of the UH problem in the data on the KH estimators, since those estimators have no mean to control for this problem.

4.4.2 Sample Statistics

There is a big divergence between the number of job-vacancies and the number of job-seekers in the sample. The total number of job-seekers and the total number of job-vacancies, which are the sum of the stock and flow of each variable, are shown in Chart 4.1. The lines in the plots show the weekly total number of job-seekers, U (black), and job-vacancies, V (red), in each local sub-area separately. The horizontal axis shows the date in weeks. The number of job-seekers has an upward trend and is growing substantially every year with very high seasonal variations, and in particular there a very large increase in the number of job-seekers in June-September each year. The number of job-seekers is affected by the schools academic year start and end dates, where the largest increase is associated with the summer break period each year.

Compared with the total number of job-seekers, the total number of the job vacancies, which is the demand side, has no similar seasonal trend. Instead the number of vacancies has an upward trend in the first half of the sample followed by a downward trend in the second half, with a concave shape inside the sample study period and is similar in all sub-areas. The trend of the total number of vacancies during the sample period indicates that they are more affected by business cycle trends than school start and end dates. However, the total number of job-vacancies counts only youth vacancies and the influence of the business cycle on the labour-market of young individuals may not be similar to the business cycle influence on the adult labour-market. The peak in the number of job-vacancies is reached in the period around the first half of 1990, this is seen clearly in West Lancashire sub-area where the majority of job-vacancies are observed.

The summary statistics in Table 4.1 shows the average duration of job-vacancies and the total number of job-vacancies during the sample period. The total number of job-vacancies is 14154 vacancies. There is an average duration of 5.76 weeks for the vacancy to get matched with a job-seeker but double the average, 10.65 weeks, for the right censored vacancy. The value of the overall average duration for all job-vacancies is closer to the average of the censored vacancies due to the high number of censored vacancies. The total number of employers in the sample is 4121, but some employers operate in more than one area. The number of job-vacancies and the duration of matched vacancies is lower in East Lancashire sub-area and is the highest in West Lancashire sub-area.

Taking the ratio of the matched vacancies in each sub-area, by dividing the row of Matched Vacancies in the first panel in the table by first row, shows that Central Lancashire is the area with the highest rate of matched vacancies, 21.4%, followed by West Lancashire then East Lancashire with percentages of matched vacancies 18.7% and 17.9% respectively. This shows that the proportion of matched vacancies in the three sub-areas is not different, but the speed at which the job-vacancies are matched, or more generally, the length of time that employers stay operating in the trading state is different between the sub-areas. Vacancies in West Lancashire sub-area have the longest average duration, followed by Central Lancashire and then East Lancashire. This order is the same for the average duration of the censored and matched vacancies combined. In West Lancashire and Central Lancashire sub-areas there are 3 job-vacancies that are matched with job-seekers after more than 2 years in the trading state in the market; there are no other late transitions. Late transitions are generally difficult to handle in duration models. They are unfavourable to the estimation of the model, and can cause many difficulties in the interpretation of the results. Spells that continue to a very long duration are created by high values of v relative to the other spells in the sample. This increases the variance of the unobserved heterogeneity in the sample and makes the estimation of the distribution parameters of the $g(v)$, $\widehat{\boldsymbol{\eta}}$, less efficient. They are sign of a high 'weeding out' problem in the data.

The number of job-vacancies in the first row in Table 4.1 is the sample size of the survival analysis data-shape, which is the data-shape that is used in kernel hazard estimation. The number in the second row, the Job Vacancies-weeks row, is the size of the pooled sample that is used in the estimation of MPH models. The size of the pooled sample depends on the duration of the individual job-vacancies, and becomes very large if some spells are too long, which creates difficulties for the estimation of MPH models. The job-vacancies that continue for a large number of weeks are more likely to be right censored and be uninformative about the true hazard rate.

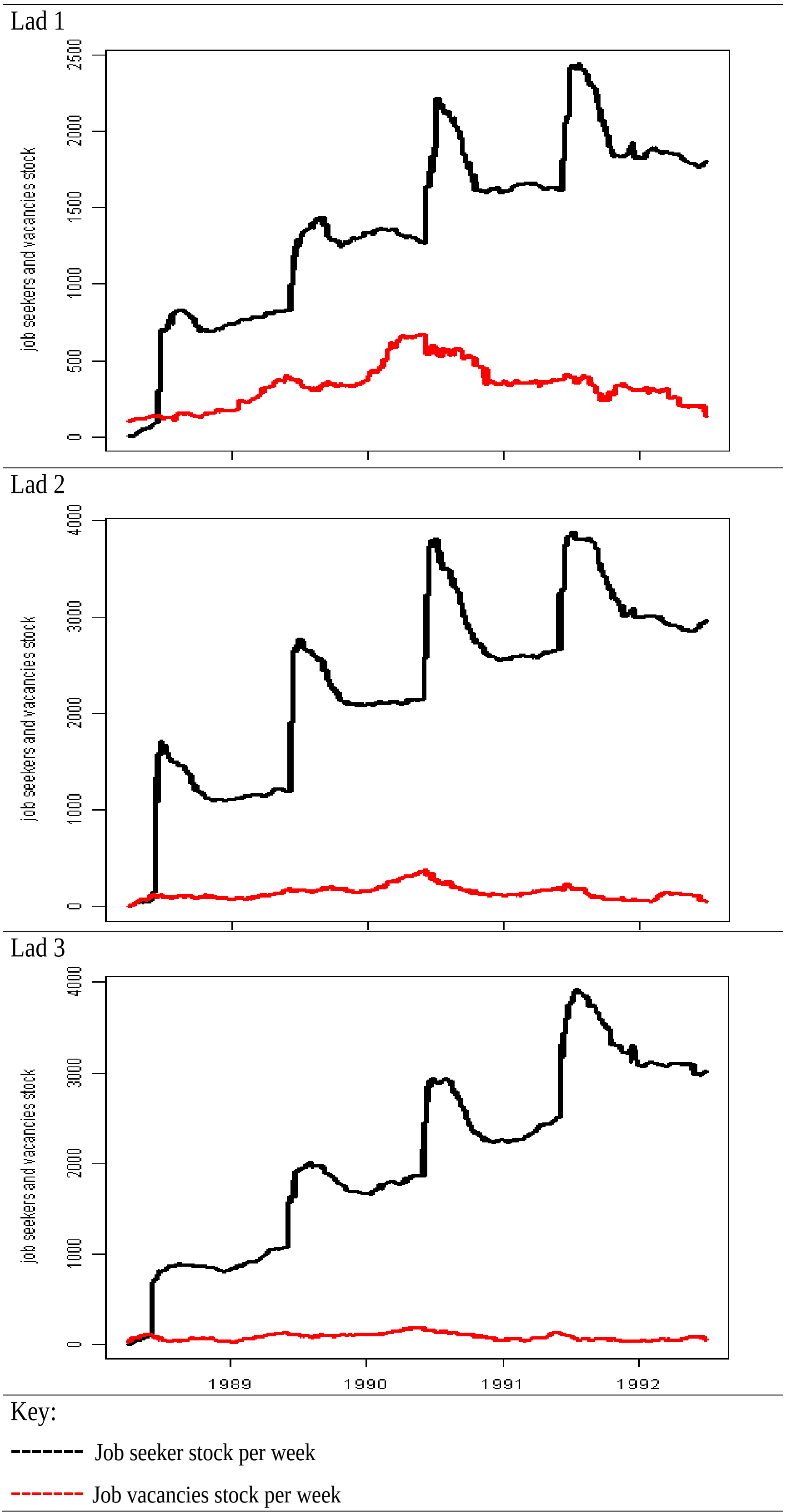


Figure 4.1: The job-seekers and the job vacancies stocks.

Table 4.1: Flow Summary Statistics[1]

	Lancashire			
	West	Central	East	All
Job Vacancies[2]	5251	5180	3723	14154
Job Vacancies-weeks[2]	79015	36128	22080	137223
Vacancy order	2739	3825	2992	9555
Employers	1212	1530	1384	4121
Matched Vacancies	983	1110	668	2761
Right Censored Vacancies	4268	4070	3055	11393
Average Duration (all)	15.05	6.98	5.94	9.70
Average Duration (censored)	16.49	7.63	6.53	10.65
Average Duration (uncensored)	8.81	4.58	3.21	5.76
Spells > 23 weeks	1075	328	146	1549
Transitions (> 23 weeks)	108	28	9	145
Censored (> 23 weeks)	967	300	137	1404
Spells > 52 weeks	309	45	39	393
Transitions (> 52 weeks)	16	8	1	25
Censored (> 52 weeks)	293	37	38	368
Spells > 104 weeks	36	4	8	48
Transitions (> 104 weeks)	2	1	0	3
Censored (> 104 weeks)	34	3	8	45

[1] The sample period is from March 1988 to June 1992.
[2] The number of job-vacancies is the sample size of flow sample in the survival analysis data shape. The number of job Vacancies-weeks is the sample size of the unbalance panel data shape (the pooled sample) that is used to estimate the (M)PH models.

4.5 The Results

The estimation results are shown for the parametric estimators first then for the non-paramtric estimators in the next two sub-sections. The estimated hazard results are delivered in a very different manner. For the parametric estimators, the output is delivered in a form of estimated coefficients for $\log U$ and $\log V$ and the dummy variables in the model. The kernel nonparametric estimators on the other hand provide the results of the estimated models in form of estimated bandwidths, $\widehat{\mathbf{h}}$, in addition to $\widehat{\gamma}_0$ in the CEKH estimator. Accordingly, generating the discussion of the results by interpreting the output of each estimation technique separately first is useful before moving to the comparison of the results.

4.5.1 Parametric Models

The coefficients of the independent variables and the variance of the unobserved heterogeneity components of the MPH models are presented in Table 4.2. For brevity, only the coefficients of the $\log U$ and $\log V$ are shown, with the estimated elasticities. The coefficients for the homogeneous PH models are reported in the first two columns and the coefficients of the MPH models with Gaussian mixture and the MPH with nonparametric mixtures are presented in columns 3 and 4, respectively. The panels in the table show the models that are estimated for the Lancashire area on the top, then the results of the models that estimated in each of the West Lancashire, Central Lancashire and East Lancashire sub-areas on the bottom. The value the coefficient of $\log j^o$ in the homogeneous models indicate a highly negative spurious duration dependence. Despite the fact that the Weibull shape parameter is not identified from the value of this coefficient, we can argue, from the evidences from the mathematical proofs in Chapter 3, that the shape parameter of the underlying continuous duration time is less than 1.

The values of elasticities of job-seekers and job-vacancies in the homogeneous models, β and α, that are estimated by the PH models are very close. The nonparametric specification of the baseline hazard improves the estimation, since the log likelihood value, $logL$, in the models with the piecewise baseline hazard is higher than in the models with Weibull baseline hazard. Therefore, the MPH models are estimated with the piecewise baseline hazards only. The mixture models on the other hand show improved log likelihood values when the distribution of the unobserved heterogeneity is relaxed from a Gaussian parametric distribution to nonparametric discrete distribution. The NP mixture is the parametric model with the most relaxed parametric assumption in the baseline hazard and the unobserved heterogeneity distribution in this research, and show the highest improvement in log Likelihood values at the same time.

The unobserved component v is defined as a random effects in the mixture models, defined as constant over time and varies among employers. This is more plausible in the random effects assumption than assuming that v varies over job-vacancies, so the vacancies that are opened by the same employer have the same value of v. Based on the sample statistics in Table 4.1, the employers in the local areas in the sample seem to have different heterogeneity. The late transitions in the sample indicate vacancies that associated with large values of v, which means that there are some employers with characteristics that are different than the others employers in the area. The employers with large value of v in our data waited longer period than other employers to have their vacancies filled by job-seekers, some employers waited more than two years, this implies very large value of v and high variance to those employer. The estimated variance of v in the NP mixture PH models is the highest in the sub-area that shows the highest number of late transitions and the lowest in the sub-area that has no late transitions, East Lancashire.

The values of the Akaike Information Criterion AIC and Bayesian Information Criterion BIC are reported for the models in each area. According to the BIC, the homogeneous models are improved after the piecewise baseline hazard is introduced, and the mixture models are improved when the distribution of the UH is relaxed to a nonparametric distribution. The value of BIC increased from 23591.86 in the homogeneous model with Weibull baseline hazard, to 23654.58 in the homogeneous models with the semi-parametric baseline hazard. In the sub-areas models BIC shows the highest improvement in the piecewise PH model in East Lancashire sub-area. For the mixture models BIC increased substantially from 22803.67 to 22911.58 in the models that estimated the Lancashire region, and from 5413.075 to 5440.929 in the models that estimated in East Lancashire sub-area. The inconsistency in the Akaike Information Criterion AIC on the other hand, is due to the regular problem of that it fails to choose the right models in many cases, see Davidson and MacKinnon (2004, Chapter 15, Section 15.4).

The variance of the unobserved components as estimated in the Gaussian mixture PH model seems as estimated from a mis-specified distribution for v, particularly in Central Lancashire sub-area, where it estimates $Var\,(v)$ less than in East Lancashire which contradict with information in Table 4.1. The results of the mixture models are different and confusing if they are critically compared with each other in the sub-areas that have large unobserved heterogeneity. If we denoted to the UH in West Lancashire, Central Lancashire and East Lancashire by v_1, v_2 and v_3 respectively, the numbers of late transitions in Table 4.1 imply that $Var(v_1) > Var(v_2) > Var(v_3)$. The results in Table 4.2 show that the Gaussian mixture estimates $\widehat{Var(v_1)} > \widehat{Var(v_3)} > \widehat{Var(v_2)}$, but NP mixture model estimates $\widehat{Var(v_1)} > \widehat{Var(v_2)} > \widehat{Var(v_3)}$.

The coefficients of the PH model that does not control for the UH are expected to be closer to zero than the coefficients that estimated by the MPH. For $\log U$ and $\log V$, the coefficients of the homogeneous PH models are closer to zero in absolute values than the coefficients that estimated by the MPH models, which indicates that the models are improved and the bias in the estimated coefficients is corrected when the MPH is used. For $log\ V$ the value of the estimated coefficient improved from -0.344 to -0.525 in the model of West Lancashire sub-area and from -0.687 to -0.793 in the model of East Lancashire sub-area. In Central Lancashire, however, this pattern is not estimated. However, the evidence from the coefficients of the Gaussian mixture model are inconsistent with the evidence from the estimated variance. It is likely that the Gaussian mixture overestimate the coefficients, estimated coefficients that have absolute value that substantially higher than zero.

With the results that are shown, our conclusion about the parametric models is that the NP mixture shows more reliable estimates relative to the Gaussian mixture and the models without the unobserved heterogeneity controlled.

Table 4.2: Coefficients of the proportional hazard models[1]

region		Weibull (homo)	Piecewise (homo)	Gaussian Mixture[2]	NP Mixture
Lancashire	$log\ j^o$	-0.794 (0.018)			
	$d_{Central}$	-0.798 (0.133)	-0.800 (0.133)	-1.366 (0.203)	-1.418 (0.196)
	d_{East}	-0.951 (0.156)	-0.959 (0.156)	-1.532 (0.235)	-1.557 (0.229)
	$log\ V$	-0.670 (0.090)	-0.676 (0.091)	-0.756 (0.135)	-0.765 (0.132)
	$log\ U(=\alpha)$	1.328 (0.154)	1.324 (0.154)	1.887 (0.223)	1.836 (0.216)
	β	0.330	0.324	0.244	0.235
	log L	-11671.72	-11655.76	-11224.39	-11213.29
	AIC[3]	23385.44	23369.53	22508.79	22508.58
	BIC[3]	23591.86	23654.58	22803.67	22911.58
	$Var\,(v)^4$			4.022	5.050
West	$log\ j^o$	-0.771 (0.027)			
	$log\ V$	-0.344 (0.276)	-0.314 (0.277)	-0.632 (0.360)	-0.525 (0.313)
	$log\ U(=\alpha)$	2.223 (0.483)	2.119 (0.478)	3.203 (0.608)	2.585 (0.529)
	β	0.656	0.686	0.368	0.475
	log L	-4652.149	-4637.604	-4319.380	-4317.146
	AIC[3]	9342.297	9329.208	8694.760	8712.291
	BIC[3]	9518.568	9579.697	8954.527	9074.109
	$Var\,(v)^4$			7.038	24.650
Central	$log\ j^o$	-0.766 (0.030)			
	$log\ V$	-0.641 (0.154)	-0.673 (0.154)	-0.622 (0.197)	-0.712 (0.209)
	$log\ U(=\alpha)$	1.666 (0.331)	1.721 (0.333)	1.673 (0.378)	1.557 (0.365)
	β	0.359	0.327	0.378	0.288
	log L	-4367.356	-4357.776	-4265.806	-4265.688
	AIC[3]	8772.713	8767.553	8587.612	8599.376
	BIC[3]	8934.115	8987.835	8825.467	8888.2
	$Var\,(v)^4$			1.991	3.042
East	$log\ j^o$	-0.902 (0.044)			
	$log\ V$	-0.689 (0.242)	-0.740 (0.243)	-0.845 (0.300)	-0.793 (0.290)
	$log\ U(=\alpha)$	1.734 (0.433)	1.789 (0.440)	2.074 (0.553)	1.930 (0.515)
	β	0.311	0.260	0.155	0.207
	log L	-2604.1992	-2597.286	-2566.503	-2565.427
	AIC[3]	5246.400	5248.572	5189.007	5192.854
	BIC[3]	5398.446	5464.637	5413.075	5440.929
	$Var\,(v)^4$			2.274	2.1153

[1] Piecewise (homo), Gaussian Mixture and NP Mixture models use step-function specification of the baseline hazard with ten grouped weeks dummies. The grouped weeks are: $[0,1)$, $[1,2)$, $[2,4)$, $[4,6)$, $[6,8)$, $[8,13)$, $[13,26)$, $[26,39)$, $[39,52)$, $[52,\infty)$. The coefficients of the steps dummies are not shown.
[2] Gaussian Mixture models are estimated using a number of quadrature points Q=24.
[3] AIC and BIC are used to determine the number of the mixture distribution mass points in each NP Mixture model. The number of the mass points that emerge in each model are: 7 mass points of the models of Lancashire and West Lancashire sub-area, 5 mass point in the model of Central Lancashire sub-area, and 3 mass points for the model in East Lancashire sub-area.
[4] $Var\,(v)$ is the variance of the unobserved heterogeneity components.

4.5.2 Kernel Hazard Models

The nonparametric kernel hazard models are estimated using the naive bandwidths as starting values to the MLCV-Risk method. Beran kernel hazard BKH and conditional external kernel hazard CEKH, show comparable values of the MLCV-Risk objective function at the optimum values of the bandwidths, even when the values of the objective function at the naïve bandwidths are extremely different. A summary of the results for the kernel hazard estimators is reported in Table 4.3. Three sets of bandwidths and objective function values in each panel are presented, which are: the naïve bandwidths, BKH optimum MLCV-Risk bandwidths and CEKH optimum MLCV-Risk bandwidths. The naïve bandwidths are not applied in any process but as starting values of the MLCV-Risk method. The hazard rates are predicted with the MLCV-Risk optimum bandwidths. Obviously, the maximum likelihood objective function shows a substantial improve for the MLCV-Risk bandwidth over the naïve bandwidths.

However, for the kernel hazards estimators in Lancashire region, the MLCV-Risk has not improved from the naïve bandwidths. This is possibly caused by the increase in the sample size and the number of independent variables in the model of the Lancashire region, giving more difficulties in the computation of the objective function. Another reason, but to the advantage of the kernel conditional hazard estimator, is that the increase in the sample size adds more consistency in the results. The inclusion of more independent variables also accommodates more observed heterogeneity. For these two reasons, it is more likely to find that the optimum bandwidths of the conditional density and the conditional hazard are close. However, it is difficult empirically to investigate whether the MLCV-Risk bandwidths are the same as the naïve bandwidths when different starting values are used, due to the long time that this new task may take.

The number of job-vacancies in the sample is 14154 vacancies, and the naïve bandwidths are achieved after more the 5 hours of processing in the HPC. It is possible to start the estimation again from other starting values, but this will be very time consuming. In addition, if the MLCV-Risk function shows progress toward some optimum bandwidths, they are more likely to be some bandwidths that are close in values to the naïve bandwidths. Looking at the $\log L$ values of the models in Table 4.2 and the values of the objective function in Table 4.3, where the original formula of the likelihood function is the same, gives an insight on the expected improve that the MLCV-Risk bandwidths will make to the hazard over the naïve bandwidths. In the model in the Lancashire region the objective function at the naïve bandwidths is already substantially higher than all the values of $\log L$ in the first panel in Table 4.2. In the Piecewise (homo) model $\log L = -11655.76$ and in the NP mixture $\log L = -11213.29$, compared with the values of the objective function of the kernel hazard estimators -10724.97 and -10570.03 for the BKH and CEKH respectively, which are higher than the values in the parametric models. Comparing this with the improvement that is achieved in the models in the sub-areas supports our claim that the naïve bandwidths and the MLCV-Risk bandwidths are possibly too close and that it is not necessary to pay the extra computation cost to estimate the MLCV-Risk bandwidths.

For the models in the sub-areas the naïve bandwidths of $log\ U$ and $log\ V$ are substantially lower than the MLCV-Risk bandwidths in the CEKH and BKH estimators, which indicates that the naïve bandwidths over-smooth the hazard rates. The bandwidth of the duration times on the other hand, are estimated lower by the MLCV-Risk. The values of the bandwidths are not very informative about the quality of the estimated kernel hazard, but the substantial improve in the value of the log likelihood function (the objective function of the MLCV-Risk) illustrates that the kernel hazard estimator are better off using the MLCV-Risk bandwidths than the naïve bandwidths. The values of the MLCV-Risk bandwidths do not show strange behaviour except in BKH in Central Lancashire, where the bandwidth of $\log V$ is estimated very large.

In spite of this limited information about the significance of the independent variables on the hazard rate, the information in Table 4.3 show nice convergence relative to the amount of the numerical computation that involved in the estimation. It also shows good quality for the kernel estimators since the values of the optimum bandwidths do not indicate an oversmoothing or an undersmoothing of the independent variables in the model. This also illustrates the fact that all the variables that are included in the model are relevant variables.

The effect of the unobserved heterogeneity in the kernel conditional hazard estimators is examined by comparing the MLCV-Risk bandwidths that are estimated in the local sub-areas. The magnitude of the bandwidths of the BKH and CEKH estimators is similar to the magnitude of the coefficient and variance v that are estimated in the MHP models in Table 4.2. In the area with the highest unobserved heterogeneity, West Lancashire sub-area, the MLCV-Risk bandwidths of the BKH and CEKH estimators are extremely different, while for the sub-area with the lowest unobserved heterogeneity, East Lancashire, the MLCV-Risk bandwidths are substantially less different. Studying the effect of UH on kernel hazard estimators, theoretically and empirically, is attractive, as this results imply that kernel estimators are affected differently.

In the areas with high unobserved heterogeneity, West and Central Lancashire sub-areas, CEKH perform better than BKH in term of the values of the estimated bandwidths. CEKH estimator, in contrast to the BKH estimator smooths the discrete time transition variable. Our intuition is that the performance of CEKH estimator overcome that of BKH estimator for that CEKH is smooths more variables, which may imply that smoothing makes the estimator less affected by the UH problem as Froelich (2006) claims. This result is encouraging to do more research in the effect of UK in kernel estimators.

Finally, the bandwidth of the $dmonth$ variable is well estimated in all models. The seasonal variations are detected in the kernel hazard rate estimators effectively by the CKEH and BKH estimators. The only variable with the MLCV-Risk bandwidth value that estimated tolerantly in all models is the $dmonth$ variable.

Table 4.3: Conditional kernel hazard bandwidths[1]

Region		Naive[1]	Beran Kernel Hazard	External Kernel Hazard
Lancashire	j^o	0.211500		0.211500
	$log\ U$	1.078088	1.078088	1.078088
	$log\ V$	0.015727	0.015727	0.015727
	$dmonth$	0.814899	0.814899	0.814899
	lad	0.590069	0.590069	0.590069
	Obj. fn (naïve)[2]		-10724.97	-10570.03
	Obj. fn (MLCV-Risk)[3]		-10724.97	-10570.03
	Execution time (h)	5.2915	216.1317	122.3389
West	j^o	0.267892		0.301522
	$log\ U$	0.030051	0.472083	0.297836
	$log\ V$	0.016243	0.276469	0.048854
	$dmonth$	0.338863	0.515488	0.460587
	Obj. fn (naïve)[2]		-6888.39	-4225.92
	Obj. fn (MLCV-Risk)[3]		-3981.01	-3945.90
	Execution time (h)	0.2997	174.0182	104.1565
Central	j^o	0.147044		0.007090
	$log\ U$	0.394634	2.516729	2.831348
	$log\ V$	0.078130	24.556184	1.551502
	$dmonth$	0.739984	0.506349	0.488774
	Obj. fn (naïve)[2]		-4296.03	-4235.03
	Obj. fn (MLCV-Risk)[3]		-3899.00	-3893.60
	Execution time (h)	0.3431	97.3131	117.0946
East	j^o	0.097844		0.051905
	$log\ U$	0.248588	1.265907	1.284935
	$log\ V$	0.187331	0.411490	0.381787
	$dmonth$	0.842971	0.670388	0.672655
	Obj. fn (naïve)[2]		-2352.29	-2383.78
	Obj. fn (MLCV-Risk)[3]		-2310.14	-2342.07
	Execution time (h)	0.1350	25.8423	19.7866

[1] Kernel hazard estimators do not control for the unobserved heterogeneity problem
[2] The naïve bandwidths are estimated using the MLCV method for conditional density, the starting values are the normal-reference bandwidths for the continuous variables and 0.5 for the categorical variables.
[3] The MLCV-Risk bandwidths are estimated by maximizing the objective function in (4.35), the naïve bandwidths are used as starting values.

4.5.3 The Comparison of the Results

A graphical presentation is used to illustrate the differences in the results and compare the predictions of the proportional hazard models and kernel estimators. Figures 4.2 to 4.7 show plots of the predicted conditional hazard of the piecewise homogeneous PH model, Gaussian mixture and NP mixture PH models, in addition to the predictions of the conditional KH estimators. The months July and August of each of the year 1989, 1990 and 1991 are presented in the first three figures, they are the months of the highest increase in the flow of young job-seekers to the labour-market and thus the duration of the vacancies is expected to be the lowest during this period of the year.

The months February and March are presented in the second three figures. In contrast to the months in the first three figures, February and March are the months with a low flow of young job-seekers, most of the job-seekers are in the stock during this period. The hazard that estimated by each estimation technique are predicted up to the 52^{th} week and at the weighted mean of U and V using the local sub-areas to weight the variables. The plots present the results of the sample of Lancashire region where more heterogeneity is explained in the model.

The slope of the predictions of the piecewise homogeneous PH (the black solid step line) is more negative than the prediction of the MPH models (the coloured solid step lines) in all plots, where spurious duration dependence is predicted by the homogeneous model since it is not controlling for UH. The spurious duration dependence underestimates the increase in the hazard rate at the later weeks in the spell. The plots show good performance of the kernel hazard estimators (the dashed step lines) compared with the performance of the parametric models. The kernel hazard predictions are closer to the piecewise homogeneous PH model since both do not control for the effect of the unobserved heterogeneity. The predictions from the parametric and the nonparametric estimators are closer in the years 1989 and 1990 in the plots than in the year 1991. This is due to the captured effect of the $year$ and $month$ dummies in the parametric models compared with the captured effect of the $dmonth$ categorical variable in the nonparametric kernel hazard estimators. The difference between the predictions of the estimation technique increases in the later months in the flow sample period, since these are the months with high values of $dmonth$. The kernel hazard predictions are relatively unchanged in value in the later months in the flow sample period. This is particularly predicted in the first week in the state but there are some changes in the second week and beyond. The (M)PH predictions are substantially different in value and slope throughout the spell length. This demonstrates that the two estimation techniques pick up different effects of the flow sample months variable in the model.

The seasonal variations in the job matching function are generated by the seasonal flows of the job-seeker in the labour-market. The effect of the weekly number of U and V on the hazard rate of the job-vacancies is the effect that the year and month variables aim to capture. We illustrate this effect using the theory of the job matching function of Pissarides (1990), and the line graphs in Figure 4.1. The number of job-seekers creates the seasonal variation in the hazard rate, so that the higher the difference between the number of job-seekers and number of job-vacancies, with $U \gg V$, the higher the average number of matches per week, m, and the higher hazard rate. The trends of U and V indicate an increasing differences in the flow sample duration, which means an increasing hazard rate. The ceteris paribus effect of the labour-market tightness, ρ, on the hazard of the job-vacancies is increasing the hazard with the decrease in ρ. So, the hazard should be with an upward slope. Similarly, the months of the high shifts in U are expected to be coincide with high shifts in the predicted hazard rate. The plots in Figures 4.2 to 4.7 show the opposite manner for the predictions of the (M)PH estimators, where they indicate that the hazard rate is decreasing in the later months in the flow sample. Also they indicate that the months with the large differences on the number of job-seekers and job-vacancies are characterised by lower hazard rates, which violates a clear proposition of the job matching models theory.

Figures 4.8 and 4.9 compare between the ability of the proportional hazard method and the kernel hazard method to estimate the seasonal differences in the predictions. The horizontal axis in the plots presents the discrete ordered values of the elapsed months in the flow sample. Each discrete ordered value coincides with a date in months and years in the flow sample period. The horizontal axis in months in Figures 4.8 and 4.9 covers the same period as that covered in the horizontal axis in Figure 4.1 in weeks. Figure 4.8 shows the estimated hazard rate in the first week only for each sub-areas, and Figure 4.9 shows the estimated hazard for the second week in the same sub-areas. The kernel hazard estimators show remarkable performance in estimating hazard rate that captures the effects of the difference between U and V, i.e. the labour-market tightness, in the model. The increase in the number of job-seekers appears to increase the hazard rate in the second week, in Figure 4.9, more than the increase in the first week. This is obvious in the model, where the matching date in the sample is the date when the matched job-seeker started to work in the position, which is more likely to be in the week after the announcement of the job-vacancy in the LCS if the contact took place in the first week. Generally, the kernel hazard estimator gives hazard rates that strongly agree with the effect of the job market tightness and the labour-market matching theory.

The last comparison aims to investigate whether the difference in the estimated seasonal effect of U is the same in the proportional hazard models, if the year and month dummies are replaced by 51 dummy variables that represent the discrete ordered variable $dmonth$ that used in the kernel estimator. This step aims to repeat the same method of comparison that used in Chapter 2, trying to making the difference between the results of the parametric and the nonparametric kernel estimators less by enriching the parametric model with coefficients. The new PH model is estimated without controlling for the unobserved heterogeneity in the sample, where the plots show that the MPH models performed worse than the homogeneous PH model in capturing the effect of the difference between U and V. The new model has 36 more coefficients than the piecewise model in the discussion above.

The predicted hazard for the new PH model is compared with the kernel hazard in Figure 4.10. The coefficient rich PH model is still unable to estimate the effect of the labour market tightness. Using a semi-parametric specification for U and V in the cloglog model to add more coefficients and slope dummies is problematic, because of the wide range of U and V values. The variables U and V need to be grouped first before converting them to dummy variables. This requires defining of a number of bins and bin widths of the grouped variables, which is an arbitrary process that depends on the experience more than on statistical rules. Additionally, for U and V variables, the grouping will smooth out useful information inside the bins, and the slope coefficients will not correctly estimate the effects in the model. Then we conclude that trying to enrich the parametric PH models with coefficients to reduce the difference between them and the kernel hazard estimators is not an effective technique.

4.6 Conclusions

This research presents an empirical example for the performance of the conditional kernel hazard estimators that are developed in Chapter 4. The example uses a flow sample from Lancashire Careers Service data covers the period from March 1988 to June 1992. The research tries to estimate a hazard rate for job-vacancies, which is less considered in the literature than the estimation of hazard of unemployment spells. The kernel hazard estimators are compared with the (mixed) proportional hazard estimators using a step-function (piecewise) specification of the baseline hazard and two choices of the distribution of the unobserved heterogeneity, a Gaussian distribution and a nonparametric discrete distribution. The kernel hazard estimators are relatively underdeveloped, but despite this they show satisfactory performance, which makes them attractive for future research. The MLCV-Risk method is used to estimate the bandwidths of the KH estimators, they show an outstanding performance compared with the naïve bandwidths.

The kernel hazard estimators do not control for the unobserved heterogeneity, unlike the MPH estimators, and elasticities of the labour demand and supply in the matching function are not directly produced in the model. The variables in the model have to be smoothed with kernel function derivatives to estimate the elasticities. The total number of job-seekers and the total number of job-vacancies in the kernel estimators are taken at the last week that is observed for the vacancy in the trading state in the sample. The kernel hazard models use the survival analysis data shape, while the (M)PH model use the unbalanced panel data shape, and U and V are observed weekly during the spell for the vacancy in the trading state. The comparison between the techniques is restricted and weakened due to that the KH estimators do not use all the information in the sample like the (M)PH estimator.

The estimated duration dependence that is shown in the Figures 4.2 to 4.7 illustrates that the hazard rate is closely predicted by the PH and the kernel hazard estimators in the early months in the flow sample period, but the predictions are substantially different in the later months. This is due to that the kernel hazard estimators are more able to pick up the effect of labour-market tightness than the PH and the MPH estimators. At the later months in the flow sample period, the job market tightness is extremely low, and leads to an increase in the hazard rate. The relationship is predicted by the kernel hazard estimators correctly, whiles the proportional hazard estimator predicts the reverse effect. The (M)PH estimators have completely failed to capture this effect. The PH model shows no improve when the number of coefficients is increased in the model.

The kernel hazard estimators show that the hazard rate of job-vacancies at the end of the flow sample observation period is three times more than its value at the beginning of the period. The increase is generated by the increase in the total number of job-seekers which increased the labour-market tightness. However, this result needs more investigation with time varying regressors and with both stock and flow of job-seekers and job-vacancies included in the model as in Andrews et al. (2011).

The research compares the kernel estimators in a difficult and advanced empirical example, in which the job matching function is highly dependent on time varying independent variables. The flow sample has large unobserved heterogeneity that affects the average duration of the job-vacancies. In spite of all those factors, the kernel hazard estimator captured features in the model more than the PH models, and delivers predictions that are more consistent with the theory of job matching function and labour-market tightness than the (M)PH model. As in the previous chapter, in this chapter the comparison is not balanced between the parametric method and the kernel method. The parametric model has given the best chance that available to make them work effectively. The kernel estimators, due to their recent development and their high computation cost, are given just a moderate chance. Despite this unfair procedure, kernel hazard estimators show outstandingly good performance compared with the PH estimators, which indicate that the estimators are well constructed.

The research suggests to do more work to improve the discrete time KH estimators, so that they become powerful estimation tools in econometrics. The current kernel hazard estimators are attractive to be examined empirically in a model that include time invariant independent variables only, and a data set that does not include high unobserved heterogeneity. This will increase the opportunity for the kernel estimators to preform better than their performance in this research. The predicted duration dependence and observed heterogeneity could then be studied better than in the models that include time varying regressors.

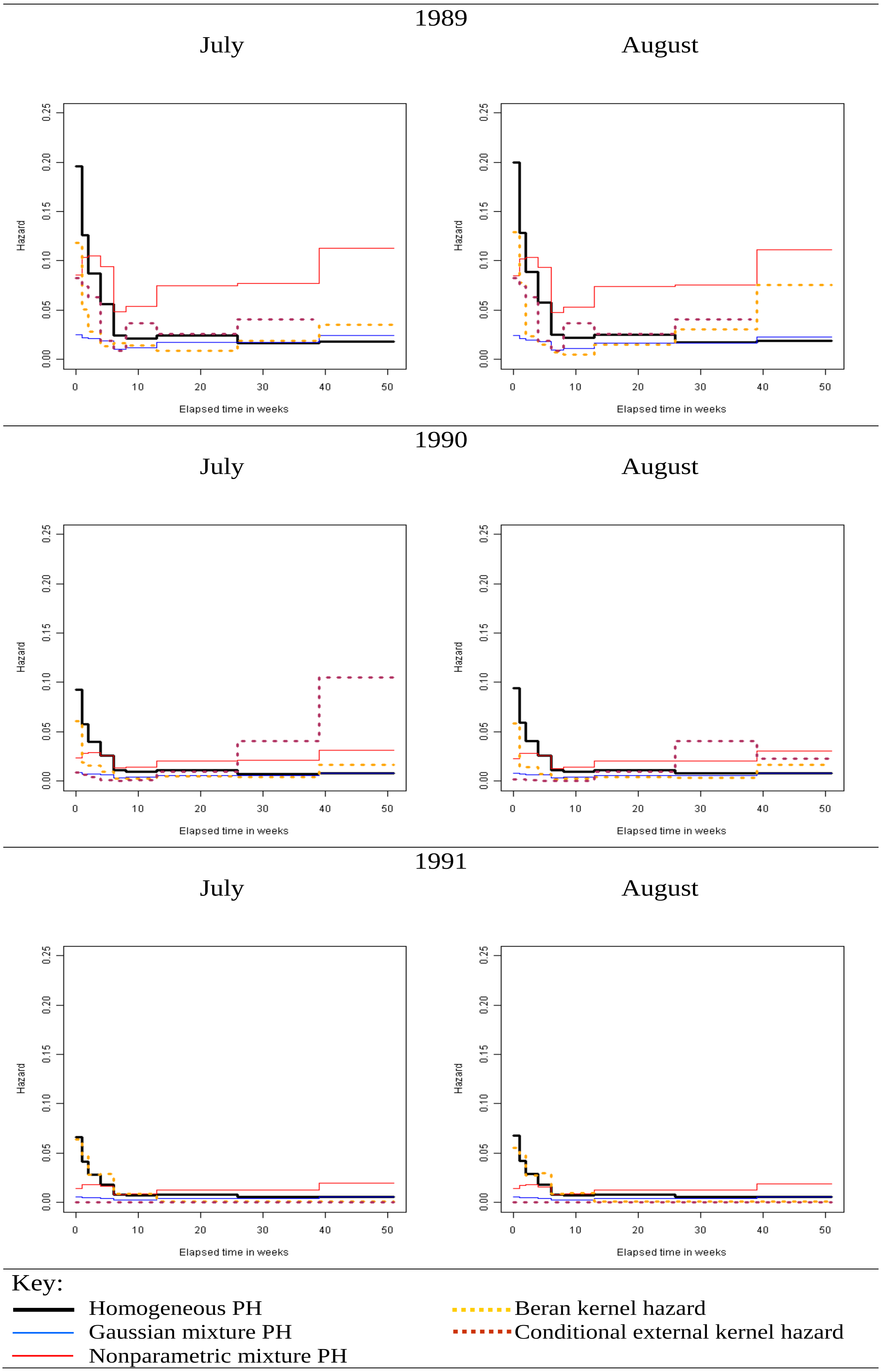


Figure 4.2: The predicted hazards in July and August months in West Lancashire sub -area.

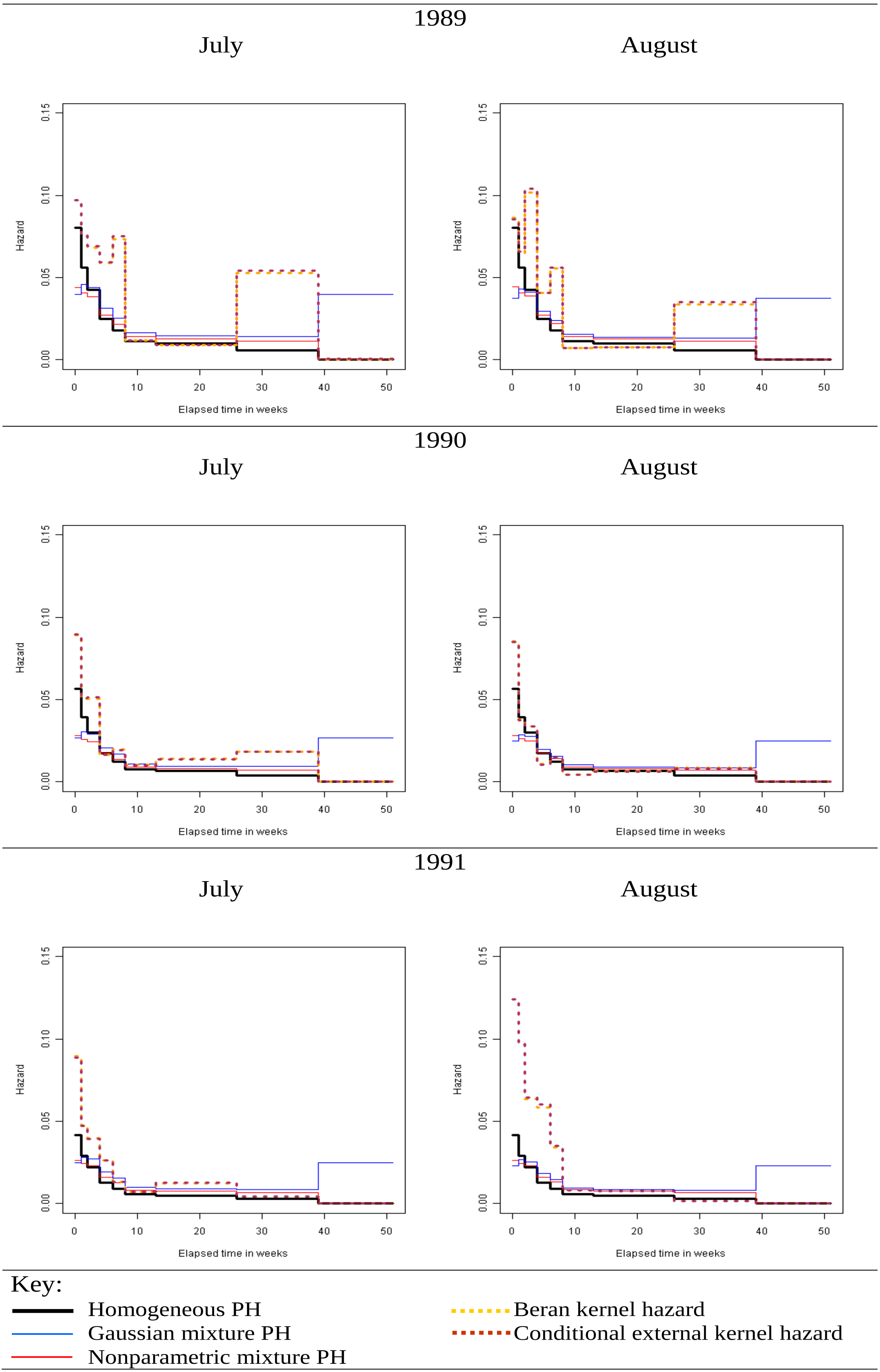


Figure 4.3: The predicted hazards in July and August months in Central Lancashire sub -area.

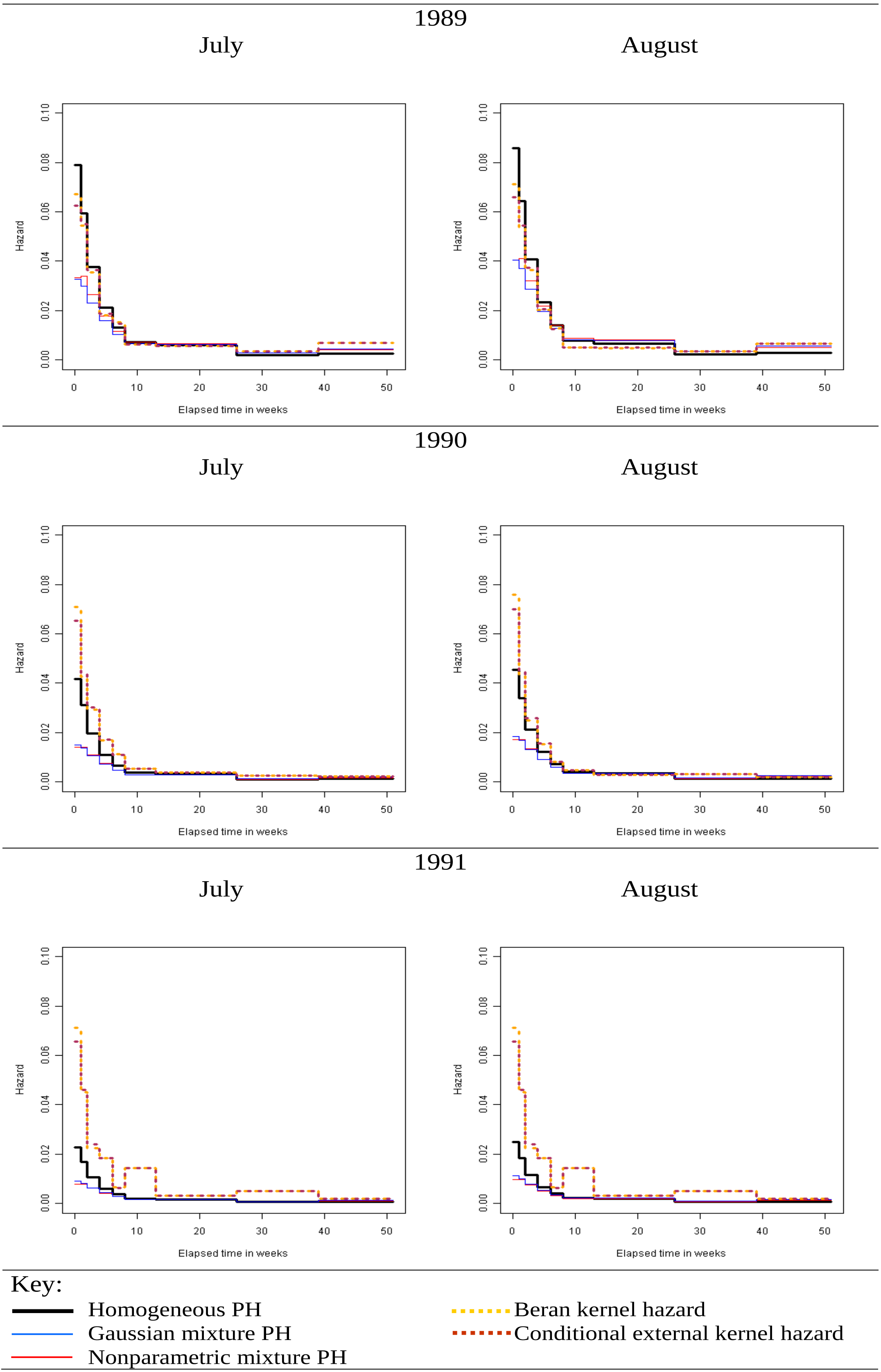


Figure 4.4: The predicted hazards in July and August months in East Lancashire sub -area.

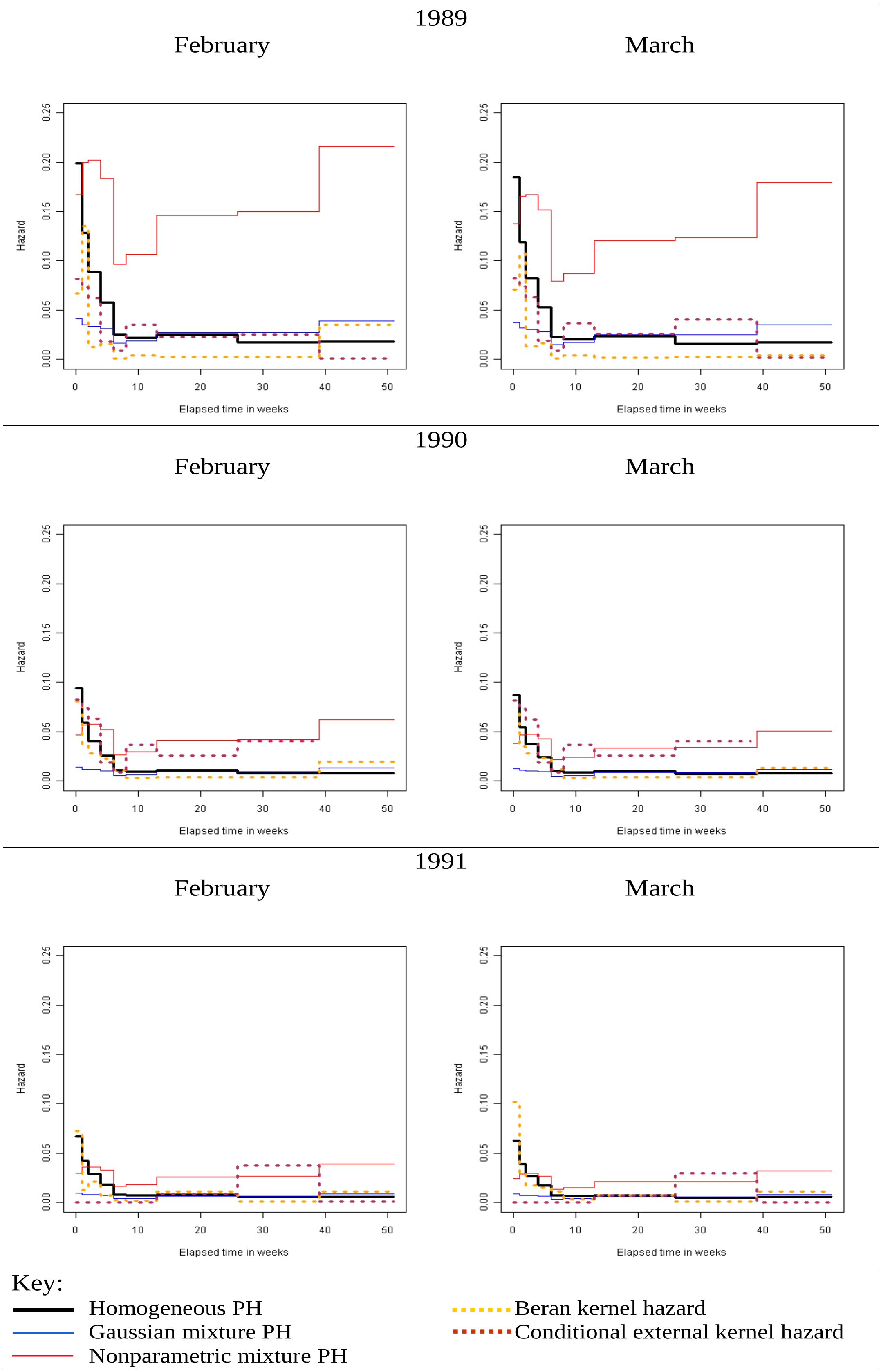


Figure 4.5: The predicted hazards in February and March months in West Lancashire sub -area.

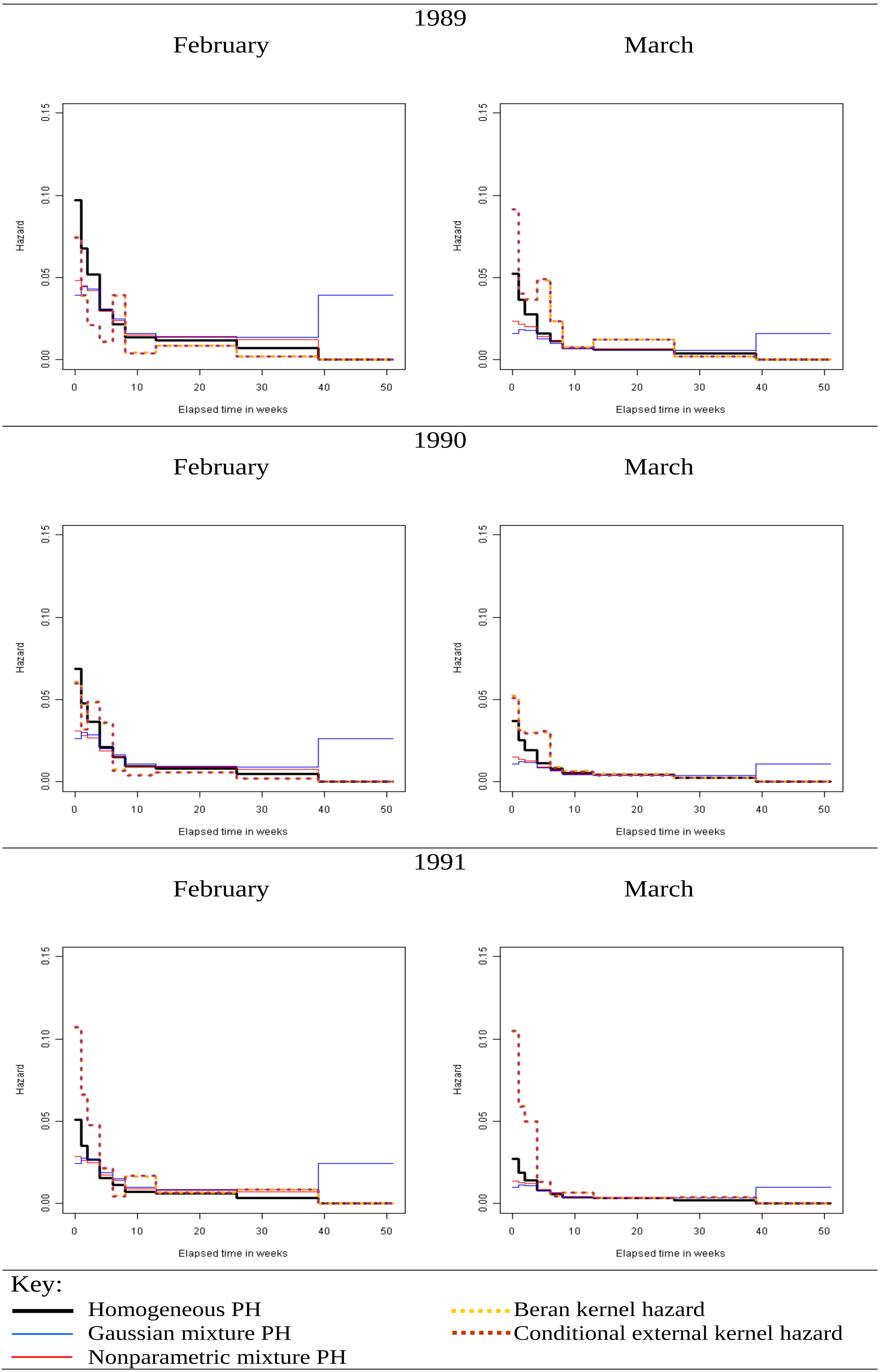


Figure 4.6: The predicted hazards in February and March months in Central Lancashire sub -area.

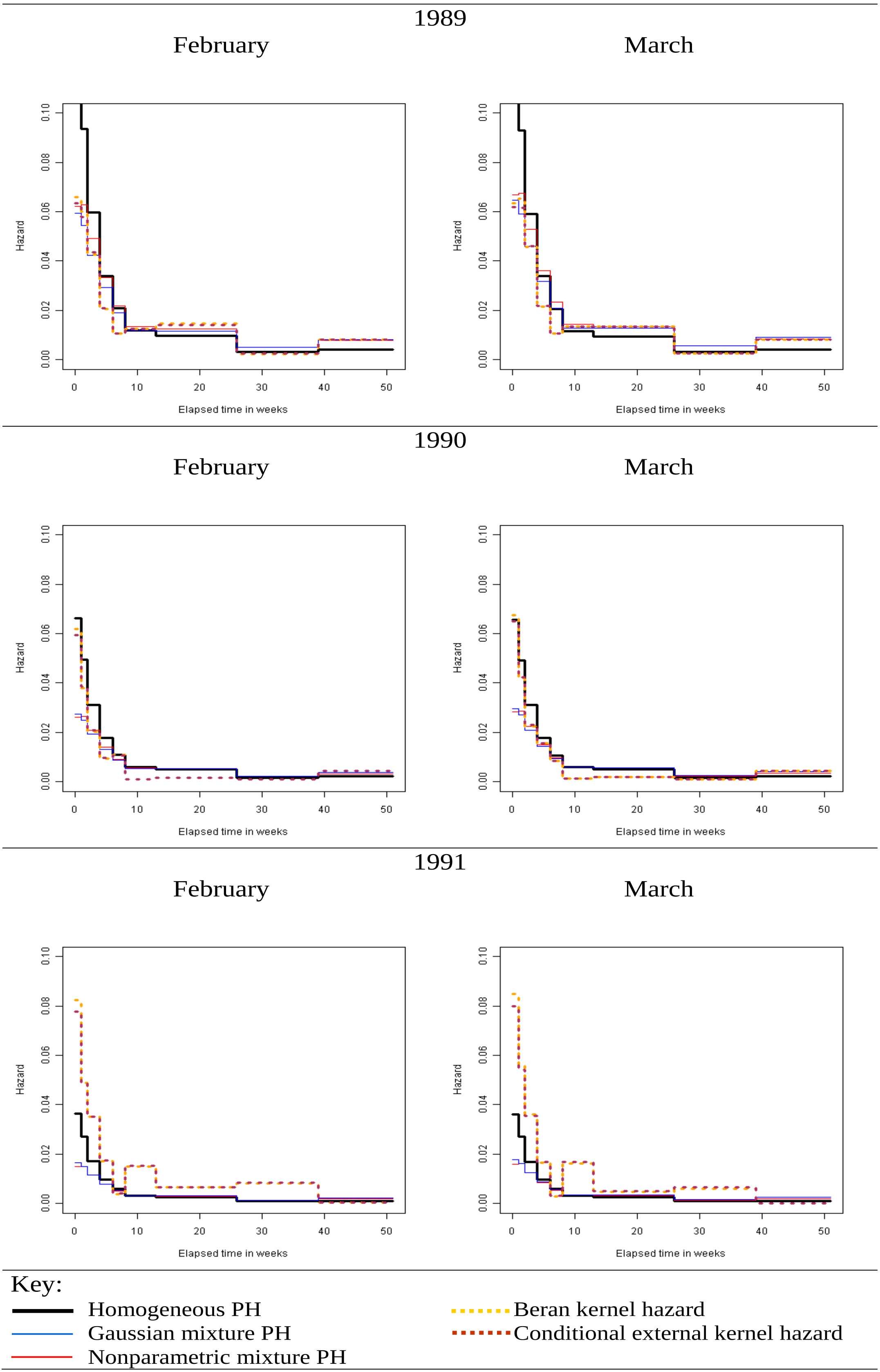


Figure 4.7: The predicted hazards in February and March months in East Lancashire sub -area.

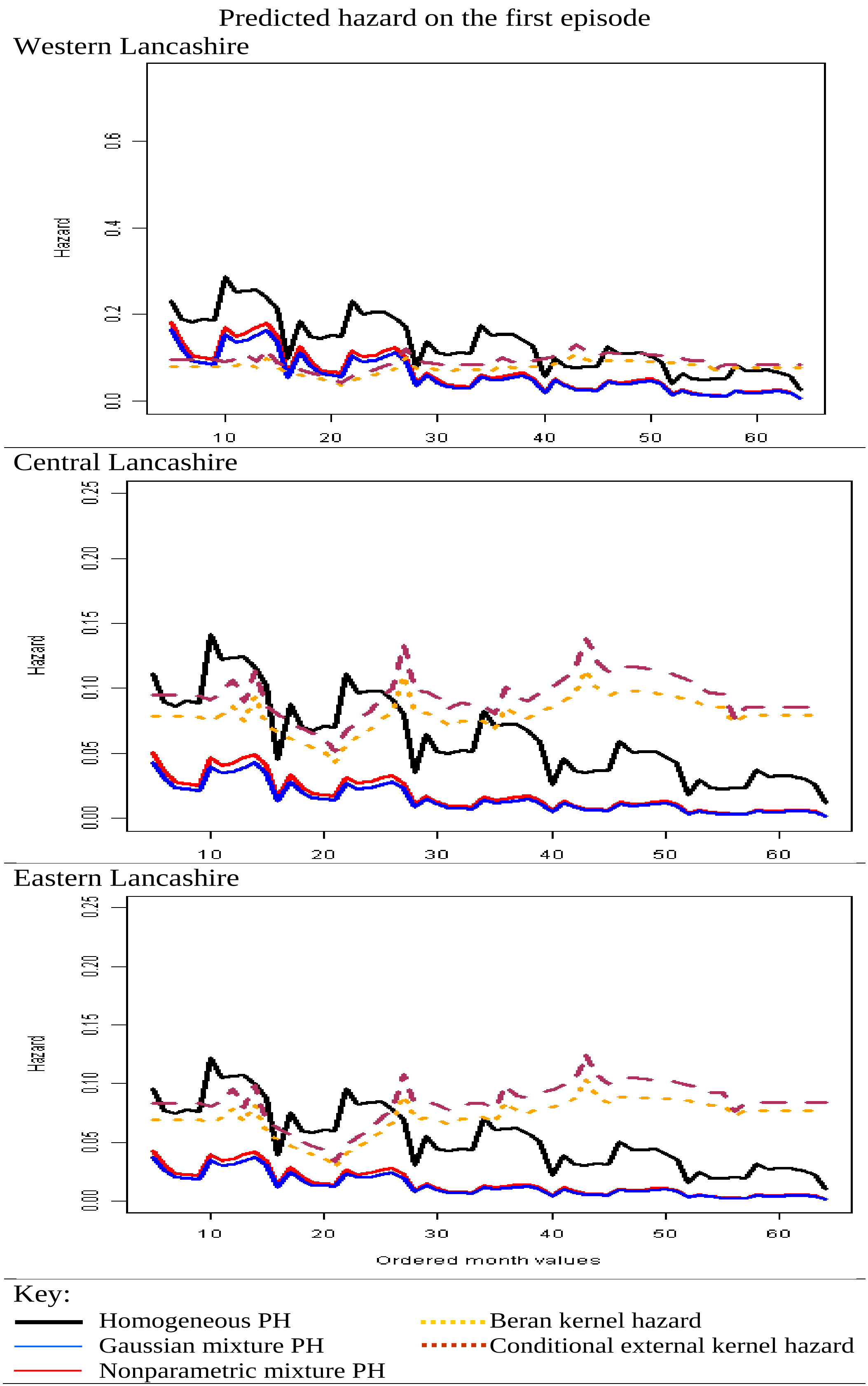


Figure 4.8: The predicted hazard rate on the first week in the trading state using the dummy variables of month and year variables in the PH model.

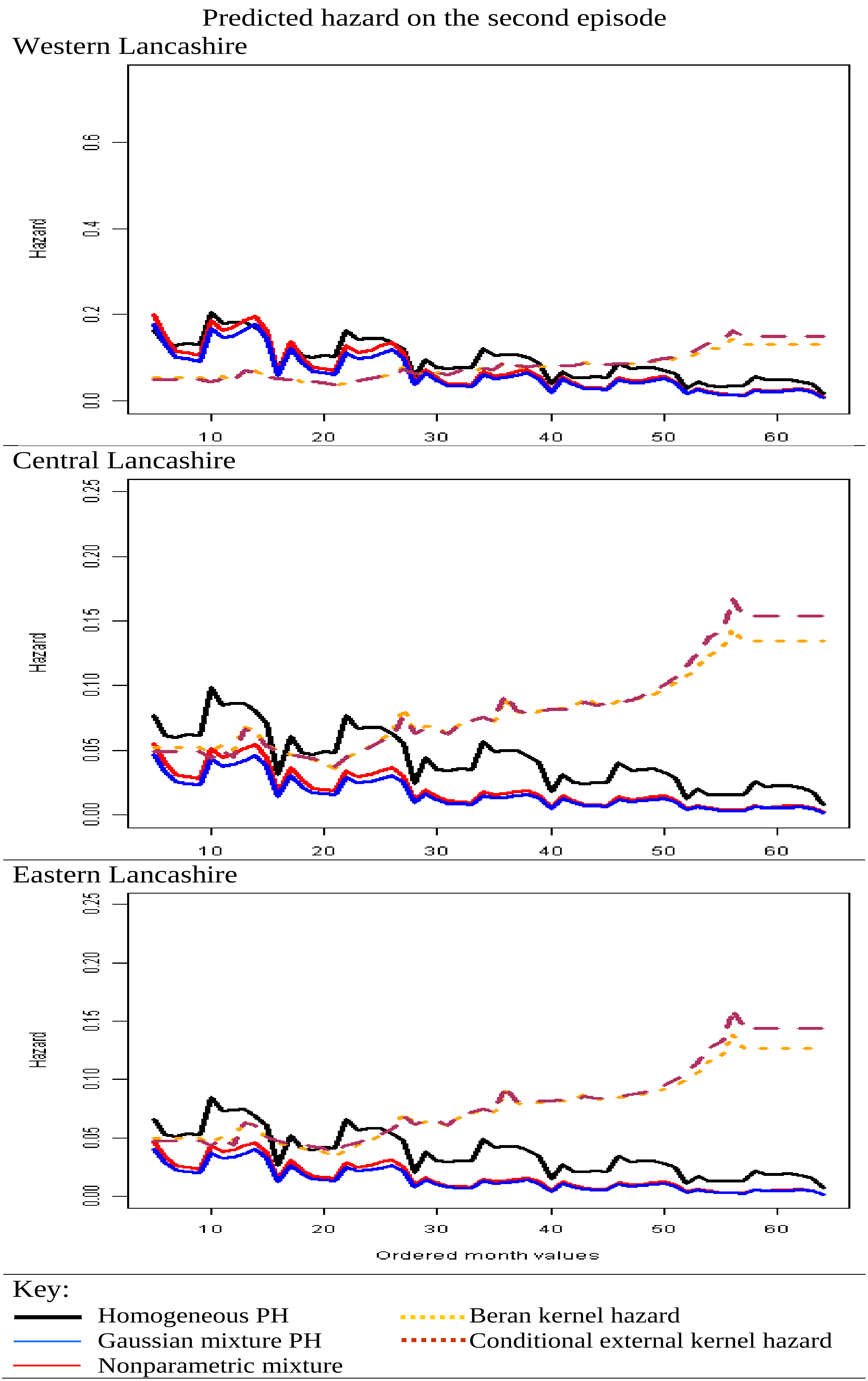


Figure 4.9: The predicted hazard rate on the second week in the trading state using the dummy variables of month and year variables in the PH model.

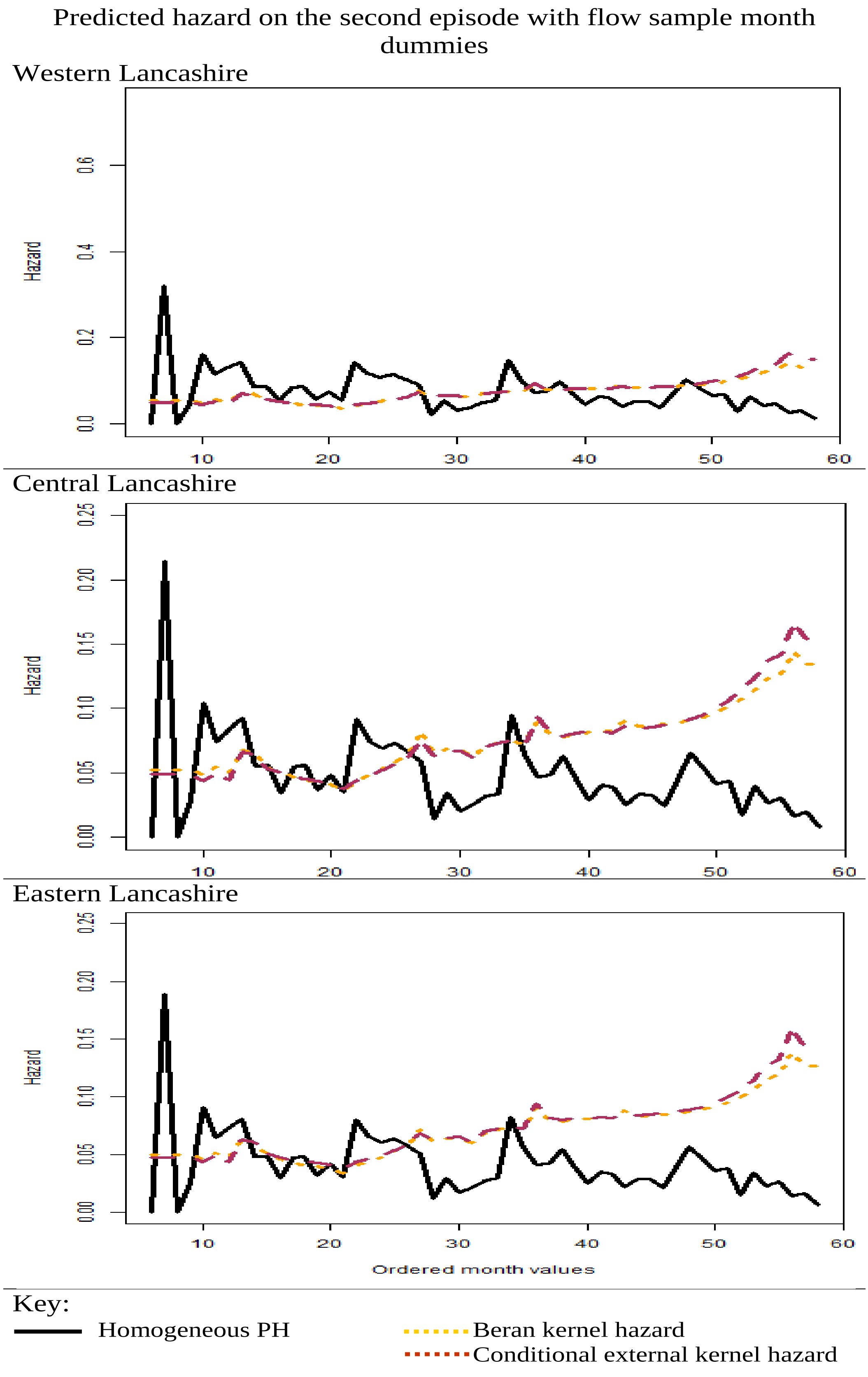


Figure 4.10: The predicted hazard rate on the second week week in the trading state using the dummy variables of the discrete ordered elapsed month in the flow sample period in the PH model (51 dummy variables).

Chapter 5

Conclusions of the Thesis

This PhD thesis examines the performance of kernel methods with fixed bandwidths (KMWFB) in estimating discrete conditional functions (DCF) conditioning on mixed data types (MDT). The discrete models that are considered in the thesis are respectively a multinomial conditional density function model and a discrete time single state duration model. We developed Chapter 2 a discrete time kernel hazard (KH) estimator and a discrete time conditional KH estimator, referred to as discrete time external kernel hazard (EKH) and discrete time conditional external kernel external hazard (CEKH) estimators respectively. The (C)EKH estimator is examined in a simulation study that is designed under the assumptions of the proportional hazard (PH) framework. Both EKH and CEKH estimator and the data generating process in the simulation study are original contributions of this thesis.

The discrete time external kernel hazard estimator is derived from the existing external kernel hazard estimator for continuous time transitions. The new discrete time hazard estimator has an advantage over the continuous time external kernel hazard estimator in that it does not smooth negative values in the lower bound in the spell, accordingly it does not have the source of the Boundary Bias. This advantage is gained from the Geometric kernel function that is used to smooth the grouped duration time in the hazard estimator. The discrete time conditional external kernel hazard estimator might be the first that being suggested in econometrics, it could be useful alternative to the exciting discrete time proportional hazard estimator in economic duration models. The simulation study of the hazard estimators is developed from existing simulation studies in Survival Models in medical and applied sciences, but we extend the designs by introducing the grouping scheme of the transition time and make the hazard in the simulation study be estimatable in both continuous and discrete times. The work of the development of the discrete time external hazard estimator and the Monte Carlo simulation study is presented in Chapter 2.

The discrete choice estimation technique that used in Chapter 2, and the parametric discrete time proportional hazard that used in in Chapter 3, are parametric estimation methods that are difficult to be specified functionally correct. The restrictive assumptions that are required for parametrically estimate the models make the produced results that do not match with the economic theory or valid only under very narrow conditions.

The multinomial conditional density function is estimated parametrically using the discrete choice method, in which the random utility theory (RUT) is the cornerstone of the development of the model. The discrete time conditional hazard is estimated parametrically using the proportional hazard (PH) estimation framework, with an underlying assumption of an existence of a baseline hazard that is equal for all individuals. In addition, it is assumed that the contribution of the independent variables is proportional, and shifts the baseline hazard either upward or downward. In both functions the restrictive assumptions lead to results that fail to capture many features in the data and contradict with the underlying economic theory.

Parametric models show limited improvement when their restrictive functional form is relaxed. However, the improvement of the parametric model requires introducing many dummy variables and interaction terms, which creates a semi-parametric function in the model. This tends to make the maximisation of the likelihood function of the parametric models difficult, and may make the model exhibit problems like multicollinearity and increase the singularity of the data matrix. This research shows empirically how difficult it is to reach a correctly specified parametric functional form in econometrics, and recommends the use of nonparametric estimation more than parametric estimation in empirical research if it is available. The computation difficulties that we faced in this work are temporary, since developments in computing systems and technologies will make nonparametric estimation easier in the near future.

5.1 Mixed Data Types Smoothing Framework

Accordingly, the role of the discrete kernel function of the dependent variable in the DCF estimation is no different to that of the link function in parametric estimation. The MDT smoothing method ultimately is the kernel framework that allows of the estimation of DCF functions. Accordingly, the kernel function that is assigned to the discrete dependent variable in the estimator is crucial for making a correct specification in the nonparametric kernel estimation approach. The kernel function of the dependent variable has to consider the assumptions and the type of the variable in the model, similarly to the link function in the parametric discrete choice estimation methods. However, the discrete kernel function that smooths the discrete dependent variable in the kernel estimators enforces very few assumptions in the model that do not restrict the interpretation of the results.

Generally kernel estimators depend on the correct choice for the kernel functions of the variables in the model. Correct kernel estimator depends on the correct choice of the kernel functions of the variables in the model, i.e to choose continuous kernel functions for the continuous variables, discrete unordered kernel functions for the discrete unordered variables and discrete ordered kernel functions for the discrete ordered variables. This property in the kernel nonparametric estimation is considered carefully in the development of the discrete time EKH and CEKH estimators in Chapter 3, where the Geometric kernel function that is used to smooth the grouped duration time in the hazard estimator is developed under assumptions that are consistent with the grouping scheme and the transition process. The mixed data types kernel estimators are not sensitive to the choice of the continuous kernel function. However, in the mixed data types kernel conditional function estimators a wrong choice of the discrete kernel function, either for the discrete dependent variable or discrete independent variables, leads to severe consequences to the results. Where this leads to inappropriately locally weighting the observations of the discrete variables and produce wrong results. The kernel method produces better results if no extreme values or outliers exist in the sample. This feature is shown in Chapter 2 for the independent variables, where the kernel conditional density estimator provides better results for all interaction terms, which is the feature that improves the capability of the MDT to effectively estimate the DCF.

In the kernel estimation of the conditional density function the MDT smoothing technique is used in both the joint and the marginal kernel density estimators. The MDT smoothing method that applied in the estimated joint density function weights the interactions of the dependent discrete variable and the discrete independent variables. The conditional density estimator is the ratio between the estimated joint density function and the estimated marginal density function. The joint behaviour of the discrete dependent variable with the discrete independent variables that captured in the estimated joint density function gives the conditional density estimator the ability to explain more heterogeneity. Which explains the source of the outstanding results that produced by the conditional density kernel estimator in Chapter 2. For the CEKH estimator in Chapters 4 the factor that improved the quality of the predicted hazard rate is the same.

In the MDT estimation framework and the DCF kernel estimation the method to estimate the bandwidths of the estimator has to be considered carefully. Bandwidth approximation methods are not supported in the MDT smoothing framework. In Chapter 3 the simulation results show that the naïve bandwidths failed to capture the effect of the discrete variables in the model, whereas the MLCV-Risk method in Chapter 4 captured the effect.

5.2 Remarks on the Results of the Empirical Models

The empirical examples that are applied in this thesis illustrate a number of facts regarding parametric and kernel nonparametric estimation for discrete conditional function. For the parametric estimation method the thesis demonstrates the following:

First, the limit dependent variable parametric estimation method assumes high homogeneity among the observations in the sample. This restricts the interpretation of the results, as shown in Chapter 2, where the MNL model estimates the same marginal effects for the categorical independent variables on the propensity to work for females in all ages. The enforced high homogeneity assumption have also led to an inaccurate results that are completely inconsistent with the economic theory, as shown in Chapter 4. Where the proportional hazard model, which uses the cloglog link function, underestimates the contribution of year and month categorical variables to the baseline hazard, which make the model fail to capture the effect of the job-market tightness.

Referring to the theory in duration models, the observations with long duration in the spell have higher variance of the unobserved components, and increase the variance of the duration time. But the frequency of such observations is low in the sample. Then the homogeneity among individuals who stay short period in the spell is higher than the homogeneity for all individuals in the sample. In parametric estimation of the hazard rate the proportional hazard model produce estimates that are biased as a result of failing of counting for the overall heterogeneity that is generated by the observations with long spells. More specifically, the estimated coefficients in the parametric PH model are better off when the sample is homogeneous, but when the heterogeneity increases, the coefficients become biased. In the kernel estimator this problem is treated automatically by the smoothing method of the MDT framework. The smoothing of the sub-survival functions (sub-distribution function) in the conditional kernel hazard estimator comprises the information regarding the individuals with long spells in the sample and their characteristic, better than the PH estimator.

Second, parametric estimation methods of DCF are easy to handle and offer many alternatives specifications for the link functions and the mixture models. However, our work shows that there is a cost regarding this advantage. The level of ambiguity on using the alternative specifications of the link functions and the mixture distribution is high. In Chapter 2 our work uses the multinomial logit (and the binomial logit) link function only because using other link functions like the multinomial probit and cloglog needs using more complicated procedures to estimate the pointwise standard errors of the predictions. Our intuition is that there may not be large differences between the predicted probabilities with those link functions and the predicted probabilities that are estimated by the logit link function. However, attempting to make the comparison in Chapter 2 between the kernel conditional density and the multinomial porbit for example, requires more work in the parametric model to estimate the pointwise standard errors and compare the precision of the predicted probabilities than that done with the logit link function.

In contrast to the choices of the link functions, the alternative specifications of the mixture distribution make large diversity in the results. This is shown clearly in Chapter 4. The definition of the mixture distribution is usually arbitrary and highly restricted, in term of the number of coefficients or the number of the mass points. Our work shows that it is likely to have inconsistent results when the assumption of mixture distribution is changed. Generally, these mentioned remarks are the reason that the estimated results of the discrete choice models are often unsatisfactory for econometricians in term of their consistency with the economic theory.

For the kernel method with fixed bandwidth the research illustrates the following points when an estimation of a DCF is attempted:

First, kernel estimation techniques are generally difficult to handle particularly when discrete variables are included in the model. The estimated results of the discrete conditional functions are distributed into cells that are created by the categories in the dependent variable in addition to the cell that are created by the discrete independent variables. The study of the results with discrete dependent variable is more difficult than the study of the results with the continuous dependent variable, due to the high number of cells that are included in the comparison. A plan to investigate the results based on experience and the sample is needed, as well as a careful design of a method of presentation of the results that depends on the research objectives. But in the presentation of the results there is high risk to the research objectives, where some important features may become difficult to show clearly in the MDT kernel estimates.

Second, it is very difficult to test the significance of the independent variables in the model using the kernel method. The performing of models selection criteria is even more difficult, due to the hard calculation of the kernel estimators. Nonparametric DCF cannot stand by their own in econometric research. Researches try to overcome this problem by using parametric version to examine the significance of the variables in the model. This process might not necessary make the significant variables that chosen in the parametric model be distinguished as relevant variables in the nonparametric estimator. The parametric estimation in our research shows low quality fits compared with the kernel nonparametric estimation, which makes it difficult to define the best fitting parametric model. The argument that says that the best fitting parametric model has to be found first is vague. Our research shows that a good fitting parametric model in very difficult to achieve, so the best fitting parametric model is a relative based on the researcher understanding of the results.

Third, Nonparametric kernel methods work better when only a few variables are included in the model due to the computation difficulties. This may imply low capability, particularly that researchers usually prefer to include many independent variables in parametric models in an attempt to explain more heterogeneity. Adding to that the curse of dimensionality problem in the nonparametric models make researcher conscious about the choice of the variables in the estimator. However, researchers have to consider the ability of kernel estimator to explain more heterogeneity on the contrary of the conventional trend of parametric estimators to enforce homogeneity in the model. In the estimation of the DCF models, the information that the nonparametric estimators explain in the model is more than that the parametric link function model can explain. Having few independent variables in the MDT estimation framework is not a disadvantage. Researchers have to note the reshaping of the independent variables to input them in the estimator of each technique. The reshaping, in the MDT framework, is more likely to make the number of covariates in the kernel estimator less than the number of covariates in the discrete dependent variable parametric model.

The research provides new evidence to the labour economic research in UK. It presents new features in female labour force participation in Chapter 2 and in the youth job market in Chapter 4, both could be useful in future research in those fields. The MDT kernel estimation framework is a effective technique to estimated DCF in econometrics, that need to be handled carefully in empirical research.

5.3 Suggestions of Further Research

The main contribution to knowledge that this thesis provides is that it applies some recently developed nonparametric techniques empirically. It focusses on the discrete conditional functions because there is general dis-satisfaction about the results of parametric discrete choice models. The novel contribution of the thesis, however, is the development of (C)EKH estimators. These estimators open new insights into approaches to estimate duration models in economics. Kernel hazard estimators for discrete time models enjoy many advantages over their corresponding estimators in continuous time, as has been illustrated in Chapter 3. This makes research in this field attractive and fruitful. The kernel nonparametric estimators can only be applied, so far, to a few types of models in econometrics. Because kernel method is weak on the estimation of models with instrumental variable or random effects and for many existing estimators the estimation of elasticities is complicated. However, with the MDT techniques and the discrete time KH estimators that we suggest in this thesis, a new field of research in kernel estimation techniques is opened.

The multinomial kernel conditional density estimator that used in this research is a simple, special case, of kernel conditional density functions in the MDT framework. The kernel conditional density estimator can be used for a multivariate conditional density, for example a conditional density with two discrete dependent variables, one could be unordered and the other could be ordered. There are potentially many interesting applications for such conditional function estimates.

The development of the discrete time kernel hazard estimator is the biggest contribution of this research. But this contribution just scratches the surface of what can be developed with the kernel method in the field of economic duration models. The MDT estimation framework provides a good basis for developing kernel estimators for competing risks models.

The effect of unobserved heterogeneity of kernel conditional density estimator and kernel conditional hazard is very attractive topic of research. An examination of this property in the kernel conditional density function estimator is interesting. The effect of the unobserved heterogeneity in the discrete choice models is considered in the literature of the parametric estimation of limit dependent variable models in econometrics. The research suggests a study to compare between the MDT conditional density estimator and the discrete choice estimation method, similar to that presented in Chapter 2, with the unobserved heterogeneity exist in the data, to examine the effect of this problem in each estimation technique. Monte Carlo simulation method that examine the effect of unobserved heterogeneity in the MDT kernel density estimator can also be included in the study.

References

Abbring, J. and Van Den Berg, G. (2007). The unobserved heterogeneity distribution in duration analysis, *Biometrika* **94**(1): 87–99.

Abdous, B., Berlinet, A. and Hengartner, N. (2003). A general theory for kernel estimation of smooth functionals of the distribution function and their derivatives, *Revue Roumaine de Mathématiques Pures et Appliquées* **48**(3): 217–232.

Ahamada, I., Flachaire, E. and Clark, A. (2011). *Non-Parametric Econometrics*, Practical Econometrics, Oxford University Press.
URL: *http://books.google.co.uk/books?id=WW2Q7nXC3iMC*

Ahmad, I. and Cerrito, P. (1994). Nonparametric estimation of joint discrete-continuous probability densities with applications, *Journal of Statistical Planning and Inference* **41**(3): 349–364.

Ahmad, I. and Lin, P. (1984). Fitting a multiple regression function, *Journal of Statistical Planning and Inference* **9**(2): 163–176.

Ahmad, I. and Ran, I. (2004). Kernel contrasts: a data-based method of choosing smoothing parameters in nonparametric density estimation, *Journal of Nonparametric Statistics* **16**(5): 671–707.

Aitchison, J. and Aitken, C. (1976). Multivariate binary discrimination by the kernel method, *Biometrika* **63**(3): 413–420.

Aitken, C. (1983). Kernel methods for the estimation of discrete distributions, *Journal of Statistical Computation and Simulation* **16**(3-4): 189–200.

Akritas, M. (2004). Nonparametric survival analysis, *Statistical Science* **19**(4): 615–623.

Allison, P. (1982). Discrete-time methods for the analysis of event histories, *Sociological methodology* **13**(1): 61–98.

Altman, N. and Leger, C. (1995). Bandwidth selection for kernel distribution function estimation, *Journal of statistical planning and inference* **46**(2): 195–214.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators, *Journal of the Royal Statistical Society. Series B (Methodological)* **32**(2): pp. 283–301.

Anderson, G. (1969). A comparison of probability density estimates, *presented IMS Annual Meeting*.

Andrews, D. (1997). A conditional kolmogorov test, *Econometrica: Journal of the Econometric Society* pp. 1097–1128.

Andrews, M., Bradley, S. and Stott, D. (2001). The school-to-work transition, skill preferences and matching.

Andrews, M., Bradley, S., Stott, D. and Upward, R. (2011). Estimating the stock-flow matching model using micro data, *Technical report*.

Andrews, M., Bradley, S. and Upward, R. (2001). Estimating the probability of a match using microeconomic data for the youth labour market, *Labour Economics* **8**(3): 335–357.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method, *Biometrika* **68**(1): 326–328.

Azzalini, A., Bowman, A. and Härdle, W. (1989). On the use of nonparametric regression for model checking, *Biometrika* **76**(1): 1–11.

Bagkavos, D. (2003). *Bias reduction in nonparametric hazard rate estimation*, PhD thesis, University of Birmingham.

Baker, M. and Melino, A. (2000). Duration dependence and nonparametric heterogeneity: A monte carlo study, *Journal of Econometrics* **96**(2): 357–393.

Baltagi, B. and Baltagi, B. (2008). *A Companion to Theoretical Econometrics*, Wiley-Blackwell.

Barnett, W., Powell, J. and Tauchen, G. (1991). *Nonparametric and semiparametric methods in econometrics and statistics: proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, number 5, Cambridge Univ Pr.

Bashtannyk, D. and Hyndman, R. (2001). Bandwidth selection for kernel conditional density estimation, *Computational Statistics & Data Analysis* **36**(3): 279–298.

Benedetti, J. (1977). On the nonparametric estimation of regression functions, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 248–253.

Beran, R. (1981). Nonparametric regression with randomly censored data, *Unpublished report, University of California, Berkeley, Department of Statistics. Downloadable from http://www.stat.ucdavis.edu/ beran/npreg.pdf* .

Bierens, H. (1987). Kernel estimators of regression functions, *Advances in Econometrics: Fifth world congress*, Vol. 1, Cambridge University Press New York, pp. 99–144.

Bingley, P. and Walker, I. (1997). The labour supply, unemployment and participation of lone mothers in in-work transfer programmes*, *The Economic Journal* **107**(444): 1375–1390.

Bingley, P. and Walker, I. (2001). Household unemployment and the labour supply of married women, *Economica* **68**(270): 157–186.

Bingley, P., Zhu, Y. and Walker, I. (2005). Education, work and wages in the uk, *German Economic Review* **6**(3): 395–414.

Bors, A. and Nasios, N. (2009). Kernel bandwidth estimation for nonparametric modeling, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* **39**(6): 1543–1555.

Botev, Z. and Kroese, D. (2008). Non-asymptotic bandwidth selection for density estimation of discrete data, *Methodology and Computing in Applied Probability* **10**(3): 435–451.

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika* **71**(2): 353–360.

Bowman, A., Hall, P. and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions, *Biometrika* **85**(4): 799–808.

Bratti, M. (2003). Labour force participation and marital fertility of italian women: the role of education, *Journal of Population economics* **16**(3): 525–554.

Brecht, B. and Brecht, L. (1996). Time-discrete nonparametric hazard model using panel data, *Statistical Papers* **37**(2): 165–176.

Burman, P. (1985). A data dependent approach to density estimation, *Probability Theory and Related Fields* **69**(4): 609–628.

Cai, Z. (2002). Regression quantiles for time series, *Econometric Theory* **18**(1): 169–192.

Cai, Z., Gu, J. and Li, Q. (2009). Some recent developments on nonparametric econometrics, *Advances in Econometrics* **25**: 495.

Calhoun, C. (1994). The impact of children on the labour supply of married women: comparative estimates from european and us data, *European Journal of Population/Revue européenne de Démographie* **10**(4): 293–318.

Cameron, A. and Trivedi, P. (2005). *Microeconometrics: methods and applications*, Cambridge University Press.

Cao, R. and Jácome, M. (2004). Presmoothed kernel density estimator for censored data, *Nonparametric Statistics* **16**(1-2): 289–309.

Cao, R. and López-de Ullibarri, I. (2007). Product-type and presmoothed hazard rate estimators with censored data, *Test* **16**(2): 355–382.

Carliner, G. (1981). Female labor force participation rates for nine ethnic groups, *The Journal of Human Resources* **16**(2): 286–293.

Chen, S. (1997). Empirical likelihood-based kernel density estimation, *Australian Journal of Statistics* **39**(1): 47–56.

Chen, W., Wang, G. and Zhang, C. (2009). Bandwidth enhancement of a microstrip-line-fed printed wide-slot antenna with a fractal-shaped slot, *Antennas and Propagation, IEEE Transactions on* **57**(7): 2176–2179.

Cheng, K. and Lin, P. (1981). Nonparametric estimation of a regression function, *Probability Theory and Related Fields* **57**(2): 223–233.

Cheng, M. and Peng, L. (2002). Regression modeling for nonparametric estimation of distribution and quantile functions, *Statistica Sinica* **12**(4): 1043–1060.

Chiu, S. (1991). Bandwidth selection for kernel density estimation, *The Annals of Statistics* **19**(4): 1883–1905.

Chiu, S. (1992). An automatic bandwidth selector for kernel density estimation, *Biometrika* **79**(4): 771–782.

Chiu, S. (1996). A comparative review of bandwidth selection for kernel density estimation, *Statistica Sinica* **6**: 129–146.

Cho, J., Ishida, I. and White, H. (2011). Revisiting tests for neglected nonlinearity using artificial neural networks, *Neural computation* **23**(5): 1133–1186.

Cho, J. and White, H. (2010). Testing for unobserved heterogeneity in exponential and weibull duration models, *Journal of econometrics* **157**(2): 458–480.

Chow, Y., Geman, S. and Wu, L. (1983). Consistent cross-validated density estimation, *The Annals of Statistics* pp. 25–38.

Chu, C. and Marron, J. (1991). Choosing a kernel regression estimator, *Statistical Science* **6**(4): 404–419.

Clark, R. (1977). Non-parametric estimation of a smooth regression function, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 107–113.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American statistical association* pp. 829–836.

Colombino, U. and Tommaso, M. (1996). Is the preference for children so low, or is the price of time so high? a simultaneous model of fertility and participation in italy with cohort effects, *Labour* **10**(3): 475–495.

Davidson, R. and MacKinnon, J. (2004). *Econometric Theory and Methods*, Oxford University Press.
URL: *http://books.google.co.uk/books?id=vqkap8AwAy0C*

Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés, *Rev. Statist. Appl* **25**(3): 5–42.

del Río, A. (1996). Comparison of bandwidth selectors in nonparametric regression under dependence, *Computational statistics & data analysis* **21**(5): 563–580.

Di Tommaso, M. and Weeks, M. (2000). Decision structures and discrete choices: an application to labour market participation and fertility, *Cambridge Working Papers in Economics* .

Dikta, G., Ghorai, J. and Schmidt, C. (2005). The central limit theorem under semiparametric random censorship models, *Journal of statistical planning and inference* **127**(1): 23–51.

Duin, R. (1976). On the choice of smoothing parameters for parzen estimators of probability density functions, *Computers, IEEE Transactions on* **100**(11): 1175–1179.

Efromovich, S. (2004). Density estimation for biased data, *The Annals of Statistics* **32**(3): 1137–1161.

Eubank, R. (1988). Spline smoothing and nonparametric regression.

Ezzahrioui, M. and Ould-Said, E. (2008). Asymptotic normality of a nonparametric estimator of the conditional mode function for functional data, *Journal of Nonparametric Statistics* **20**(1): 3–18.

Falk, M. (1984). Relative deficiency of kernel type estimators of quantiles, *The Annals of Statistics* **12**(1): 261–268.

Falk, M. (1985). Asymptotic normality of the kernel quantile estimator, *The Annals of statistics* pp. 428–433.

Falk, M. (1993). Asymptotically optimal estimators of general regression functionals, *Journal of multivariate analysis* **47**(1): 59–81.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies, *The Annals of Statistics* **21**(1): 196–216.

Fan, J., Hall, P., Martin, M. and Patil, P. (1996). On local smoothing of nonparametric curve estimators, *Journal of the American Statistical Association* pp. 258–266.

Fan, J., Heckman, N., Wand, M., of North Carolina (System). Institute of Statistics, U. and of North Carolina at Chapel Hill. Dept. of Statistics, U. (1992). *Local Polynomial Kernel regression for generalized linear models and quasi-likelihood functions*, Department of Statistics,[University of North Carolina at Chapel Hill.

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems, *Biometrika* **83**(1): 189–206.

Fan, J. and Yim, T. (2004). A crossvalidation method for estimating conditional densities, *Biometrika* **91**(4): 819–834.

Fan, Y., Li, Q. and Min, I. (2006). A nonparametric bootstrap test of conditional distributions, *Econometric Theory* **22**(04): 587–613.

Fernández, A. and Rodríquez-Poo, J. (1997). Estimation and specification testing in female labor participation models: parametric and semiparametric methods, *Econometric Reviews* **16**(2): 229–247.

Fishbein, M. (1967). A behavior theory approach to the relations between beliefs about an object and the attitude toward the object, *Readings In Attitude Theory and Measurement. New York: John Wiley & Sons* pp. 389–399.

Forslund, A. and Johansson, K. (2007). Random and stock-flow models of labour market matching-swedish evidence, *Working Paper Series* .

Froelich, M. (2006). Non-parametric regression for binary dependent variables, *The Econometrics Journal* **9**(3): 511–540.

Fryer, M. (1977). A review of some non-parametric methods of density estimation, *IMA Journal of Applied Mathematics* **20**(3): 335–354.

Gannoun, A., Saracco, J. and Yu, K. (2007). Comparison of kernel estimators of conditional distribution function and quantile regression under censoring, *Statistical Modelling* **7**(4): 329–344.

Gasser, T. and Müller, H. (1979). Kernel estimation of regression functions, *Smoothing techniques for curve estimation* pp. 23–68.

Gasser, T. and Müller, H. (1984). Estimating regression functions and their derivatives by the kernel method, *Scandinavian Journal of Statistics* pp. 171–185.

Gasser, T., Muller, H. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 238–252.

Gerfin, M. (1996). Parametric and semi-parametric estimation of the binary response model of labor market participation, *Journal of Applied Econometrics* **11**(3): 321–39.

Goodwin, B. and Holt, M. (2002). Parametric and semiparametric modeling of the off-farm labor supply of agrarian households in transition bulgaria, *American Journal of Agricultural Economics* **84**(1): 184–209.

Greco, L., Pagnotta, S. et al. (2009). Density estimation by skew-normal kernels.

Greene, W. (2003). Econometric analysis.

Guilloux, A. (2011). Contributions to nonparametric statistics for survival analysis, *Laboratoire de Statistique Thorique et Applique (LSTA), downloadable from: http://www.lsta.upmc.fr/guilloux/synthese.pdf* .

Habbema, J., Hermans, J. and Remme, J. (1978). Variable kernel density estimation in discriminant analysis, *Compstat 1978: Proceedings in Computational Statistics*, Vienna: Physica-Verlag, pp. 178–185.

Habbema, J., Hermans, J. and Van den Broek, K. (1974). A stepwise discriminant analysis program using density estimation, *Compstat*, pp. 101–110.

Habbema, J., Hermans, J. and Van Der Burgt, A. (1974). Cases of doubt in allocation problems, *Biometrika* **61**(2): 313–324.

Hall, P. (1982). Cross-validation in density estimation, *Biometrika* **69**(2): 383–390.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation, *The Annals of Statistics* **11**(4): 1156–1174.

Hall, P. and Huang, L. (2002). Unimodal density estimation using kernel methods, *Statistica Sinica* **12**(4): 965–990.

Hall, P., Huang, L., Gifford, J. and Gijbels, I. (2001). Nonparametric estimation of hazard rate under the constraint of monotonicity, *Journal of Computational and Graphical Statistics* **10**(3): 592–614.

Hall, P. and Kang, K. (2005). Unimodal kernel density estimation by data sharpening, *Statistica Sinica* **15**: 73–98.

Hall, P. and Marron, J. (1987). On the amount of noise inherent in bandwidth selection for a kernel density estimator, *The Annals of Statistics* pp. 163–181.

Hall, P., Marron, J. and Park, B. (1992). Smoothed cross-validation, *Probability Theory and Related Fields* **92**(1): 1–20.

Hall, P. and Presnell, B. (1999). Density estimation under constraints, *Journal of Computational and Graphical Statistics* pp. 259–277.

Hall, P., Racine, J. and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities, *Journal of the American Statistical Association* **99**(468): 1015–1026.

Hall, P. and Robinson, A. (2009). Reducing variability of crossvalidation for smoothing-parameter choice, *Biometrika* **96**(1): 175–186.

Hall, P. and Turlach, B. (1999). Reducing bias in curve estimation by use of weights, *Computational statistics & data analysis* **30**(1): 67–86.

Hall, P., Wolff, R. and Yao, Q. (1999). Methods for estimating a conditional distribution function, *Journal of the American Statistical Association* pp. 154–163.

Ham, J. and Rea Jr, S. (1987). Unemployment insurance and male unemployment duration in canada, *Journal of Labor Economics* pp. 325–353.

Hansen, B. (2004a). Bandwidth selection for nonparametric distribution estimation, *manuscript, University of Wisconsin* .

Hansen, B. (2004b). Nonparametric estimation of smooth conditional distributions, *Unpublished paper: Department of Economics, University of Wisconsin* .

Hardie, W. and Linton, O. (1994). Applied nonparametric methods, *Handbook of econometrics* **4**: 2295–2339.

Härdle, W. (1989). Resampling for inference from curves, *Discussion Paper Serie A* .

Härdle, W. (2004). *Nonparametric and Semiparametric Models*, Springer Series in Statistics, Springer.
URL: *http://books.google.co.uk/books?id=qPCmAOS-CoMC*

Hardle, W., Hall, P. and Marron, J. (1992). Regression smoothing parameters that are not far from their optimum, *Journal of the American Statistical Association* pp. 227–233.

Hardle, W., Janssen, P. and Serfling, R. (1988). Strong uniform consistency rates for estimators of conditional functionals, *The Annals of Statistics* pp. 1428–1449.

Hardle, W., Klinke, S., Miiller, M. and Statistique, V. (1993). Applied nonparametric smoothing techniques.

Härdle, W. and Manski, C. (1993). Nonparametric and semiparametric approaches to discrete response analysis, *Journal of Econometrics* **58**(1-2): 1–2.

Hardle, W. and Marron, J. (1985). Optimal bandwidth selection in nonparametric regression function estimation, *The Annals of Statistics* **13**(4): 1465–1481.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). Nonparametric and semiparametric models.

Hastie, T. and Loader, C. (1993). Local regression: Automatic kernel carpentry, *Statistical Science* **8**(2): 120–129.

Hayfield, T. and Racine, J. (2008). Nonparametric econometrics: The np package, *Journal of Statistical Software* **27**(5): 1–32.

Hazelton, M. and Turlach, B. (2007). Reweighted kernel density estimation, *Computational statistics & data analysis* **51**(6): 3057–3069.

Heckman, J. and Singer, B. (1984a). Econometric duration analysis, *Journal of Econometrics* **24**(1-2): 63–132.

Heckman, J. and Singer, B. (1984b). The identifiability of the proportional hazard model, *The Review of Economic Studies* **51**(2): 231–241.

Heckman, J. and Singer, B. (1984c). A method for minimizing the impact of distributional assumptions in econometric models for duration data, *Econometrica: Journal of the Econometric Society* pp. 271–320.

Heineck, G. (2004). Religion, attitudes towards working mothers and wives full-time employment, *Evidence for Austria, Germany, Italy, the UK, and the USA* (39).

Herrmann, E. (1997). Local bandwidth choice in kernel regression estimation, *Journal of Computational and Graphical Statistics* pp. 35–54.

Hess, K., Serachitopol, D. and Brown, B. (1999). Hazard function estimators: a simulation study, *Statistics in medicine* **18**(22): 3075–3088.

Hess, W. and Persson, M. (2009). Survival and death in international trade-discrete-time durations of eu imports, *Working Papers* .

Hill, M. (1983). Female labor force participation in developing and developed countries–consideration of the informal sector, *The Review of Economics and Statistics* **65**(3): 459–468.

Holmes, M., Gray, A. and Isbell, C. (2007). Fast nonparametric conditional density estimation, *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI07)*.

Horne, J. and Garton, E. (2006). Likelihood cross-validation versus least squares cross-validation for choosing the smoothing parameter in kernel home-range analysis, *Journal of Wildlife Management* **70**(3): 641–648.

Horowitz, J. (1998). *Semiparametric methods in econometrics*, Vol. 131, Springer Verlag.

Horowitz, J. (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity, *Econometrica* **67**(5): 1001–1028.

Hyndman, R., Bashtannyk, D. and Grunwald, G. (1996). Estimating and visualizing conditional densities, *Journal of Computational and Graphical Statistics* pp. 315–336.

Hyndman, R. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions, *Journal of nonparametric statistics* **14**(3): 259–278.

Izenman, A. (1991). Recent developments in nonparametric density estimation, *Journal of the American Statistical Association* pp. 205–224.

Jacobsen, J., Pearce III, J. and Rosenbloom, J. (1999). The effects of childbearing on married women's labor supply and earnings: using twin births as a natural experiment, *Journal of Human Resources* pp. 449–474.

Jácome, M. and Iglesias-Pérez, M. (2008). Presmoothed estimation with left-truncated and right-censored data, *Communications in StatisticsTheory and Methods* **37**(18): 2964–2983.

Janssen, P., Swanepoel, J. and Veraverbeke, N. (2007). Modifying the kernel distribution function estimator towards reduced bias, *Statistics* **41**(2): 93–103.

Jenkins, S. (1995). Easy estimation methods for discrete-time duration models, *Oxford bulletin of Economics and Statistics* **57**(1): 129–136.

Jenkins, S. (1997). Estimation of discrete time (grouped duration data) proportional hazards models: pgmhaz, *Stata Technical Bulletin Reprints* pp. 109–121.

Jenkins, S. (2005). Survival analysis, *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK. Downloadable from http://www. iser. essex. ac. uk/teaching/degree/stephenj/ec968/pdfs/ec968lnotesv6. pdf* .

Jenkins, S. (2006). Hshaz: Stata module to estimate discrete time (grouped data) proportional hazards models, *Statistical Software Components* .

Jiang, J. and Doksum, K. (2003). On local polynomial estimation of hazard rates and their derivatives under random censoring, *Lecture Notes-Monograph Series* pp. 463–481.

Jin, Z. and Shao, Y. (1999). On kernel estimation of a multivariate distribution function, *Statistics & probability letters* **41**(2): 163–168.

Johnson, G. (1980). The theory of labour market intervention, *Economica* **47**(187): 309–329.

Johnston, J., Johnston, J., DiNardo, J. and DiNardo, J. (2007). *Econometric Methods*, McGraw-Hill economics series, McGraw-Hill.
URL: *http://books.google.co.uk/books?id=W_KdPwAACAAJ*

Jones, M. (1995). On higher order kernels, *Journaltitle of Nonparametric Statistics* **5**(2): 215–221.

Jones, M., Marron, J. and Sheather, S. (1996). A brief survey of bandwidth selection for density estimation, *Journal of the American Statistical Association* pp. 401–407.

Jones, M. and Signorini, D. (1997). A comparison of higher-order bias kernel density estimators, *Journal of the American Statistical Association* pp. 1063–1073.

Ju, G., Li, R. and Liang, Z. (2009). Nonparametric estimation of multivariate cdf with categorical and continuous data, *Advances in Econometrics* **25**: 291.

Karunamuni, R. J. and Alberts, T. (2005). On boundary correction in kernel density estimation, *Statistical Methodology* **2**(3): 191–212.

Kim, C., Kim, S., Park, M. and Lee, H. (2006). A bias reducing technique in kernel distribution function estimation, *Computational Statistics* **21**(3): 589–601.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, Statistics for Biology and Health, Springer.
URL: *http://books.google.co.uk/books?id=jS2Cy0lezJIC*

Kumar, A. (2006). Nonparametric conditional density estimation of labour force participation, *Applied Economics Letters* **13**(13): 835–841.

Kumar, T. and Markman, J. (1975). Estimation of probability density function: A monte carlo comparison of parametric and nonparametric methods, *Third World Congress of the Econometric Society, Toronto, Canada*.

Lancaster, T. (1979). Econometric methods for the duration of unemployment, *Econometrica: Journal of the Econometric Society* pp. 939–956.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*, Econometric Society Monographs, Cambridge University Press.
URL: *http://books.google.co.uk/books?id=a0B9QgAACAAJ*

Lancaster, T. and Nickell, S. (1980). The analysis of re-employment probabilities for the unemployed, *Journal of the Royal Statistical Society. Series A (General)* pp. 141–165.

Läuter, H. and Liero, H. (2006). Nonparametric estimation and testing in survival models, *Probability, statistics and modelling in public health* pp. 319–331.

Lee, M. (1996). *Methods of moments and semiparametric econometrics for limited dependent variable models*, Springer Verlag.

Lee, T. and Solo, V. (1999). Bandwidth selection for local linear regression: a simulation study, *Computational Statistics* **14**(4): 515–532.

Lee, Y. and Lee, E. (2008). Kernel methods for estimating derivatives of conditional quantiles, *Journal of the Korean Statistical Society* **37**(4): 365–373.

Lehrer, E. (1992). The impact of children on married women's labor supply: Black-white differentials revisited, *Journal of Human Resources* pp. 422–444.

Lehrer, E. and Nerlove, M. (1986). Female labor force behavior and fertility in the united states, *Annual Review of Sociology* pp. 181–204.

Li, G. and Doss, H. (1995). An approach to nonparametric regression for life history data using local linear fitting, *The Annals of Statistics* pp. 787–823.

Li, Q., Maasoumi, E. and Racine, J. (2009). A nonparametric test for equality of distributions with mixed categorical and continuous data, *Journal of Econometrics* **148**(2): 186–200.

Li, Q. and Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data, *Journal of Multivariate Analysis* **86**(2): 266–292.

Li, Q. and Racine, J. (2004). Cross-validated local linear nonparametric regression, *Statistica Sinica* **14**(2): 485–512.

Li, Q. and Racine, J. (2007). *Nonparametric econometrics: Theory and practice*, Princeton Univ Pr.

Li, Q. and Racine, J. (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data, *Journal of Business and Economic Statistics* **26**(4): 423–434.

Li, Q., Racine, J. and Wooldridge, J. (2008). Estimating average treatment effects with continuous and discrete covariates: The case of swan-ganz catheterization, *The American Economic Review* **98**(2): 357–362.

Li, Q. and Zhou, J. (2005). The uniqueness of cross-validation selected smoothing parameters in kernel estimation of nonparametric models, *Econometric theory* pp. 1017–1025.

Liu, R. and Yang, L. (2008). Kernel estimation of multivariate cumulative distribution function, *Journal of Nonparametric Statistics* **20**(8): 661–677.

Loader, C. (1999). Bandwidth selection: classical or plug-in?, *The Annals of Statistics* **27**(2): 415–438.

Maddala, G. (1983). *Limited-dependent and qualitative variables in econometrics*, Vol. 3, Cambridge Univ Pr.

Maglad, N. and Consortium, A. E. R. (1998). *Female labour supply in Sudan*, Vol. 30, African Economic Research Consortium.

Manski, C. and McFadden, D. (1981). *Structural analysis of discrete data with econometric applications*.

Marron, J. (1994). Visual understanding of higher-order kernels, *Journal of Computational and Graphical Statistics* pp. 447–458.

Marron, J. and Padgett, W. (1987). Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples, *The Annals of Statistics* **15**(4): 1520–1535.

Mátyás, L. and Sevestre, P. (2008). *The econometrics of panel data: fundamentals and recent developments in theory and practice*, number v. 1995 in *Advanced studies in theoretical and applied econometrics*, Springer.

Mays, J. (1995). *Model robust regression: combining parametric, nonparametric, and semiparametric methods*, PhD thesis, Virginia Polytechnic Institute and State University.

Mays, J., Birch, J. and Starnes, B. (2001). Model robust regression: combining parametric, nonparametric, and semiparametric methods, *Journal of Nonparametric Statistics* **13**(2): 245–277.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior, *Frontiers in Econometrics* pp. 105–142.

McFadden, D. (1984). Econometric analysis of qualitative response models, *Handbook of econometrics* **2**: 1395–1457.

Meyer, B. (1991). Unemployment insurance and unemployment spells.

Miller, C. and Xiao, J. (1999). Effects of birth spacing and timing on mothers' labor force participation, *Atlantic economic journal* **27**(4): 410–421.

Mittelhammer, R., Mittelhammer, R., Judge, G. and Miller, D. (2000). *Econometric foundations*, Vol. 1, Cambridge Univ Pr.

Molina, I., Saei, A. and José Lombardía, M. (2007). Small area estimates of labour force participation under a multinomial logit mixed model, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170**(4): 975–1000.

Mood, A. and Graybill, F. (1963). *Introduction to the theory of statistics*, McGraw-Hill.

Muller, H. and Stadtmuller, U. (1987). Variable bandwidth kernel estimators of regression curves, *The Annals of Statistics* pp. 182–201.

Nadaraya, E. (1964). On estimating regression, *Teoriya Veroyatnostei i ee Primeneniya* **9**(1): 157–159.

Nadaraya, È. (1965). On non-parametric estimates of density functions and regression curves, *Teoriya Veroyatnostei i ee Primeneniya* **10**(1): 199–203.

Nakamura, A. and Nakamura, M. (1994). Predicting female labor supply: effects of children and recent work experience, *Journal of Human Resources* pp. 304–327.

Nicol, C. (1993). An empirical comparison of nonparametric and parametric engel functions, *Empirical Economics* **18**(2): 233–249.

Nicoletti, C. and Rondinelli, C. (2010). The (mis) specification of discrete duration models with unobserved heterogeneity: a monte carlo study, *Journal of Econometrics* **159**(1): 1–13.

Nielsen, J. and Linton, O. (1995). Kernel estimation in a nonparametric marker dependent hazard model, *The Annals of Statistics* pp. 1735–1748.

Ouyang, D., Li, Q. and Racine, J. (2006). Cross-validation and the estimation of probability distributions with categorical data, *Nonparametric Statistics* **18**(1): 69–100.

Owen, A. (2001). *Empirical likelihood*, Vol. 92, CRC press.

Pagan, A. R. and Ullah, A. (1999). *Nonparametric econometrics*, Themes in modern econometrics, Cambridge University Press, Cambridge; New York.

Park, B. and Marron, J. (1990). Comparison of data-driven bandwidth selectors, *Journal of the American Statistical Association* pp. 66–72.

Pissarides, C. (1990). Equilibrium unemployment theory.

Politis, D. and Romano, J. (1999). Multivariate density estimation with general flat-top kernels of infinite order, *Journal of Multivariate Analysis* **68**(1): 1–25.

Powell, J. and Program, P. U. E. R. (1993). *Estimation of semiparametric models*, Econometric Research Program, Princeton University.

Priestley, M. and Chao, M. (1972). Non-parametric function fitting, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 385–392.

Rabe-Hesketh, S. and Pickles, A. (2002). Multilevel selection models using gllamm, *Stata User Group Meeting. Maastricht*.

Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2001). Gllamm: A general class of multilevel models and a stata program, *Multilevel Modelling Newsletter* **13**(1): 17–23.

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*, Stata Corp.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004). Gllamm manual, *UC Berkeley Division of Biostatistics working paper series* **1160**.

Racine, J. (2008). *Nonparametric econometrics: a primer*, Vol. 3, Now Pub.

Racine, J., Hart, J. and Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models, *Econometric Reviews* **25**(4): 523–544.

Racine, J. and Hayfield, T. (2008). Package np faq, *Journal of Statistical Software* .

Racine, J., Li, Q. and Zhu, X. (2004). Kernel estimation of multivariate conditional distributions, *Annals of Economics and Finance* **5**(2): 211–235.

Racine, J., Parmeter, C. and Du, P. (2009). Constrained nonparametric kernel regression: Estimation and inference, *Technical report*, Working paper.

Rice, J. (1984). Bandwidth choice for nonparametric regression, *The Annals of Statistics* **12**(4): 1215–1230.

Rice, J. and Rosenblatt, M. (1976). Estimation of the log survivor function and hazard function, *Sankhyā: The Indian Journal of Statistics, Series A* pp. 60–78.

Robinson, P. (1988). Semiparametric econometrics: A survey, *Journal of Applied Econometrics* **3**(1): 35–51.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, *The Annals of Mathematical Statistics* pp. 832–837.

Rosenblatt, M. (1969). Conditional probability density and regression estimators, *Multivariate analysis ii* **25**: 31.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics* pp. 65–78.

Sain, S., Baggerly, K. and Scott, D. (1994). Cross-validation of multivariate densities, *Journal of the American Statistical Association* pp. 807–817.

Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions, *Journal of Statistical Planning and Inference* **35**(1): 65–75.

Scott, D. and Factor, L. (1981). Monte carlo study of three data-based nonparametric probability density estimators, *Journal of the American Statistical Association* pp. 9–15.

Scott, D. and Terrell, G. (1987). Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association* pp. 1131–1146.

Shapiro, D. and Mott, F. (1979). Labor supply behavior of prospective and new mothers, *Demography* **16**(2): 199–208.

Silverman, B. (1986). *Density estimation for statistics and data analysis*, Vol. 26, Chapman & Hall/CRC.

Simonoff, J. (1996). *Smoothing methods in statistics*, Springer Verlag.

Spanos, A. (1999). *Probability Theory and Statistical Inference: econometric modeling with observational data*, Cambridge Univ Pr.

Spierdijk, L. (2008). Nonparametric conditional hazard rate estimation: A local linear approach, *Computational Statistics & Data Analysis* **52**(5): 2419–2434.

Stone, C. (1977). Consistent nonparametric regression, *The annals of statistics* **5**(4): 595–620.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression, *The Annals of Statistics* **10**(4): 1040–1053.

Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates, *The Annals of Statistics* pp. 1285–1297.

Stute, W. (1986a). Conditional empirical processes, *The Annals of Statistics* **14**(2): 638–647.

Stute, W. (1986b). On almost sure convergence of conditional empirical distribution functions, *The Annals of Probability* **14**(3): 891–901.

Sueyoshi, G. (1995). A class of binary response models for grouped duration data, *Journal of Applied Econometrics* **10**(4): 411–431.

Tanner, M. and Wong, W. (1984). Data-based nonparametric estimation of the hazard function with applications to model diagnostis and exploratory analysis, *Journal of the american Statistical association* pp. 174–182.

Titterington, D. (1980). A comparative study of kernel-based density estimates for categorical data, *Technometrics* pp. 259–268.

Titterington, D. and Bowman, A. (1985). A comparative study of smoothing procedures for ordered categorical data, *Journal of Statistical Computation and Simulation* **21**(3-4): 291–312.

Tommaso, M. (1999). A trivariate model of participation, fertility and wages: the italian case, *Cambridge Journal of Economics* **23**(5): 623–640.

Turlach, B. (1993). Bandwidth selection in kernel density estimation: A review, *CORE and Institut de Statistique* **19**(4): 1–33.

Tutz, G. and Pritscher, L. (1996). Nonparametric estimation of discrete hazard functions, *Lifetime Data Analysis* **2**(3): 291–308.

Ullah, A. (1988a). Non-parametric estimation of econometric functionals, *Canadian Journal of Economics* pp. 625–658.

Ullah, A. (1988b). Nonparametric estimation and hypothesis testing in econometric models, *Empirical Economics* **13**(3): 223–249.

Ullah, A. (1989). *Semiparametric and nonparametric econometrics*, Studies in empirical economics, Physica-Verlag.
URL: *http://books.google.co.uk/books?id=EgmyAAAAIAAJ*

Ullah, A., Wan, A. and Chaturvedi, A. (2002). *Handbook of applied econometrics and statistical inference*, Vol. 165, Marcel Dekker Inc.

Van den Berg, G. (2001). Duration models: specification, identification and multiple durations, *Handbook of econometrics* **5**: 3381–3460.

Van Keilegom, I. and Veraverbeke, N. (2001). Hazard rate estimation in nonparametric regression with censored data, *Annals of the Institute of Statistical Mathematics* **53**(4): 730–745.

Van Ryzin, J. (1973). A histogram method of density estimation, *Communications in Statistics-Theory and Methods* **2**(6): 493–506.

Vieu, P. (1991). Nonparametric regression: optimal local bandwidth choice, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 453–464.

Wand, M. and Jones, M. (1995). *Kernel smoothing*, Vol. 60, Chapman & Hall/CRC.

Wang, B. and Wang, X. (2007). Bandwidth selection for weighted kernel density estimation, *Electronic Journal of Statistics* (0000).

Wang, M. and Van Ryzin, J. (1981). A class of smooth estimators for discrete distributions, *Biometrika* **68**(1): 301–309.

Wasserman, L. (2006). *All of nonparametric statistics*, Springer-Verlag New York Inc.

Watson, G. (1964). Smooth regression analysis, *Sankhyā: The Indian Journal of Statistics, Series A* pp. 359–372.

Watson, G. and Leadbetter, M. (1964a). Hazard analysis. i, *Biometrika* **51**(1/2): 175–184.

Watson, G. and Leadbetter, M. (1964b). Hazard analysis ii, *Sankhyā: The Indian Journal of Statistics, Series A* pp. 101–116.

Wegman, E. (1972a). Nonparametric probability density estimation, *Journal of Statistical Computation and Simulation* **1**(3): 225–245.

Wegman, E. (1972b). Nonparametric probability density estimation: I. a summary of available methods, *Technometrics* **14**(3): 533–546.

Woodroofe, M. (1970). On choosing a delta-sequence, *The Annals of Mathematical Statistics* **41**(5): 1665–1671.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*, The MIT press.

Xie, X. (1997). Children and female labour supply behaviour, *Applied Economics* **29**(10): 1303–1310.

Younes, N. and Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring, *Biometrics* pp. 1199–1211.

Yu, K. and Jones, M. (1997). A comparison of local constant and local linear regression quantile estimators, *Computational statistics & data analysis* **25**(2): 159–166.

Yu, K. and Jones, M. (1998). Local linear quantile regression, *Journal of the American Statistical Association* pp. 228–237.

Yu, K., Sun, X. and Mitra, G. (2008). Nonparametric multivariate conditional distribution and quantile regression, *Technical report*, Working paper, CARISMA, Brunel University.

Zheng, J. (2000). A consistent test of conditional parametric distributions, *Econometric Theory* **16**(5): 667–691.

Zheng, X. (2008). Testing for discrete choice models, *Economics Letters* **98**(2): 176–184.

Ziegler, K. (2002). On nonparametric kernel estimation of the mode of the regression function in the random design model, *Journal of Nonparametric Statistics* **14**(6): 749–774.

Appendix A

Key of the Notation

Variables

X^c	continuous variable
X^d	discrete variable
Y^d	discrete dependent variable (multinomial)
U^d	binary dependent variable
J^o	discrete (grouped) transition time (the episode of transition number)
Y^c	underlying continuous transition time
$Y^c_{j^o}$	underlying continuous transition time at the end of episode j
C^c	underlying continuous right censoring time
$\triangle$	right censoring indicator ($=0$ if right censored)
T^c	underlying continuous duration time

Vectors

$\boldsymbol{X}$	independent mixed variables vector
$\boldsymbol{X}^c$	sub-vector of the independent continuous variables in X
$\boldsymbol{X}^d$	sub-vector of the independent discrete categorical variables in X
$\boldsymbol{h}$	$(=[\boldsymbol{h}_{\mathbf{x}^c}, \boldsymbol{\gamma}_{\mathbf{x}^d}])$ mixed independent variables bandwidths vector
$\boldsymbol{h}_{\mathbf{x}^c}$	sub-vector of the continuous independent variable bandwidths
$\boldsymbol{\gamma}_{\mathbf{x}^d}$	sub-vector of the categorical independent variable bandwidths
$\boldsymbol{p}$	vector of length B of the probabilities for intervals of the discretised continuous time density

Scalars

b	the binwidth, interval length (the length of the episode)
B	number of grouped time intervals
h_0	continuous time variable bandwidth
γ_0	grouped time variable bandwidth
h_{x^c}	continuous independent variable bandwidth
γ_{x^d}	categorical independent variable bandwidth
λ	scale parameter of the Weibull transition time density function
k	shape parameter of the Weibull transition time density function
$\kappa^*_{(\cdot)}$	standardised value of the smoothing specification test of discrete choice models

Functions

iid	independent and identically distributed
$l(\cdot,\cdot,\cdot)$	categorical variable kernel function
$k(\cdot)$	continuous variable kernel function
$w(\cdot)$	continuous variable kernel function for independent variable
$W(\cdot,\cdot,\cdot)$	product of mixed variables kernel functions
$A(\cdot,\cdot,\cdot)$	Nadaraya-Watson weights of a product of mixed variables kernel functions
$f(t^c)$	probability density function of the continuous transition time variable at t^c (used in section 3 only)
$F(t^c)$	distribution function of the continuous transition time variable at t^c (used in section 3 only)
$S(t^c)$	survival function of the continuous transition time variable at t^c (used in section 3 only)
$r(\cdot)$	probability density function of the continuous right censoring time variable
$R(\cdot)$	distribution function of the continuous right censoring time variable
$\overline{R}(\cdot)$	survival function of the continuous right censoring time variable
$g(\cdot)$	probability density function of the continuous observed duration time variable
$G(\cdot)$	distribution function of the continuous observed duration time variable
$\overline{G}(\cdot)$	survival function of the continuous observed duration time variable
$\phi(\cdot)$	continuous time hazard rate
$f(t^c\vert\mathbf{x})$	conditional probability density function of the continuous transition time variable at t^c conditional on the mixed type independent variables $X=\mathbf{x}$ (used in section 3 only)

$F(t^c\|\mathbf{x})$	conditional distribution function of the continuous transition time variable at t^c conditional on the mixed type independent variables $X = \mathbf{x}$ (used in section 3 only)
$S(t^c\|\mathbf{x})$	conditional survival function of the continuous transition time variable at t^c conditional on the mixed type independent variables $X = \mathbf{x}$ (used in section 3 only)
$f_{Y^c}(Y^c_{j^o})$	probability density function of the continuous transition time under the grouping scheme at the end of episode j^o
$F_{Y^c}(Y^c_{j^o})$	distribution function of the continuous transition time under the grouping scheme at the end of episode j^o
$S_{Y^c}(Y^c_{j^o})$	survival function of the continuous transition time under the grouping scheme at the end of episode j^o
$f(j^o)$	or (f_{j^o}) the probability mass of the grouped time variable at episode j^o (discretised from f_{Y^c})
$F(j^o)$	or (F_{j^o}) the distribution function of the grouped time variable at the end episode j^o
$S(j^o)$	or (S_{j^o}) the survival function of the grouped time variable at the end of episode j^o
$\theta(j^o)$	the hazard probability of the grouped time at episode j^o
$J^o_{\{Q^{(\rho\%)}_{Y^c}\}}$	the interval that include the ρ^{th} quantile of the continuous time under the grouping scheme Y^c

Estimates

$\widehat{F}_n(\cdot)$	frequency estimated distribution function of the continuous time model
$\widehat{F}_{\widehat{\boldsymbol{h}},1}(\cdot)$	kernel estimated sub-distribution function of the uncensored observations in the risk set
$\widehat{F}_{\widehat{\boldsymbol{h}},2}(\cdot)$	kernel estimated sub-distribution function of the observations in the risk set
$\widehat{S}_n(\cdot)$	frequency estimated survival function of the continuous time model
$\widehat{S}_{\widehat{\boldsymbol{h}},1}(\cdot)$	kernel estimated sub-survival function of the uncensored observations in the risk set
$\widehat{S}_{\widehat{\boldsymbol{h}},2}(\cdot)$	kernel estimated sub-survival function of the observations in the risk set
$\widehat{\phi}^{DNH}(\cdot)$	unconditional density implied estimated hazard rate for the continuous time transition model
$\widehat{\phi}^{AFT}(\cdot)$	Continuous time Accelerated Failure Time regression estimated hazard rate
$\widehat{\phi}^{EKH}(\cdot)$	unconditional external kernel hazard estimated for the continuous time transition model
$\widehat{\phi}^{CDNH}(\cdot\|\mathbf{x})$	continuous time conditional kernel density implied hazard estimated conditioning on the covariates set $\mathbf{x}$
$\widehat{\phi}^{CEKH}(\cdot\|\mathbf{x})$	estimated continuous time conditional external kernel hazard when $\boldsymbol{X} = \mathbf{x}$
$\widehat{\phi}^{BKH}(\cdot\|\mathbf{x})$	estimated continuous time Beran Kernel Hazard when $\boldsymbol{X} = \mathbf{x}$
$\widehat{f}(j^o)$	kernel estimated probability mass function of the grouped time for episode j^o
$\widehat{F}(j^o)$	kernel estimated distribution function of the grouped time at the end episode j^o
$\widehat{S}(j^o)$	kernel estimated survival function of the grouped time at the end of episode j^o
$\widehat{\theta}^{DNH}(j^o)$	estimated density implied unconditional hazard of the grouped time at episode j^o
$\widehat{\theta}^{EKH}(j^o)$	estimated external kernel unconditional hazard of the grouped time at episode j^o
$\widehat{\theta}^{WPH}(j^o)$	estimated unconditional proportional hazard function using $log(j^o)$ variable for the elapsed grouped time in the spell
$\widehat{\theta}^{SPH}(j^o)$	estimated unconditional proportional hazard function with piecewise (step function) specification
$\widehat{\theta}^{CDNH}(j\|\mathbf{x})$	estimated density implied conditional hazard of the grouped time at episode j^o
$\widehat{\theta}^{CEKH}(j\|\mathbf{x})$	estimated conditional external kernel hazard of the grouped time at episode j^o and $\mathbf{x}$
$\widehat{\theta}^{BKH}(j\|\mathbf{x})$	estimated Beran kernel hazard of the grouped time at episode j conditioning on the covariates set $\mathbf{x}$
$\widehat{\theta}^{CWPH}(j\|\mathbf{x})$	estimated conditional proportional hazard function using $log(j^o)$ specification of the baseline hazard conditioning on the covariates set $\mathbf{x}$
$\widehat{\theta}^{CSPH}(j\|\mathbf{x})$	estimated piecewise conditional proportional hazard function conditioning on the covariates set $\mathbf{x}$, uses a step function (semi-parametric) specification of the baseline hazard
$\widehat{\lambda}^{(\cdot)}$	estimated scale parameter of Weibull transition time density function from model $(\cdot)$
$\widehat{k}^{(\cdot)}$	estimated shape parameter of Weibull transition time density function from model $(\cdot)$
$\widehat{\alpha}_0^{(\cdot)}$	estimated intercept from model $(\cdot)$
$\widehat{\alpha}_{log(j^o)}^{(\cdot)}$	estimated coefficient of $log(j^o)$ from a WPH or $CWPH$ model (indicated by $(\cdot)$)
$\widehat{\beta}^{\star}_{x^c}$	estimated coefficient of the continuous time independent variable x^c from a AFT
$\widehat{\beta}^{\star}_{x^d}$	estimated coefficient of the dummy independent variable associated with x^d from a AFT
$\widehat{\beta}^{(\cdot)}_{x^c}$	estimated coefficient of the continuous time independent variable x^c from a PH model (indicated by $(\cdot)$)
$\widehat{\beta}^{(\cdot)}_{x^d}$	estimated coefficient of the dummy independent variable associated with x^d from a PH model (indicated by $(\cdot)$)
$\widehat{\gamma}_0$	estimated naive bandwidth of the grouped time variable
$\widehat{\boldsymbol{h}}$	estimated naive bandwidth of the covariates of the conditional hazard
$\widehat{h}_{x^c}$	estimated naive bandwidth of the continuous covariate independent variable x^c
$\widehat{\gamma}_{x^d}$	estimated naive bandwidth of the discrete covariate independent variable x^c

Appendix B

Appendices of Chapter 1

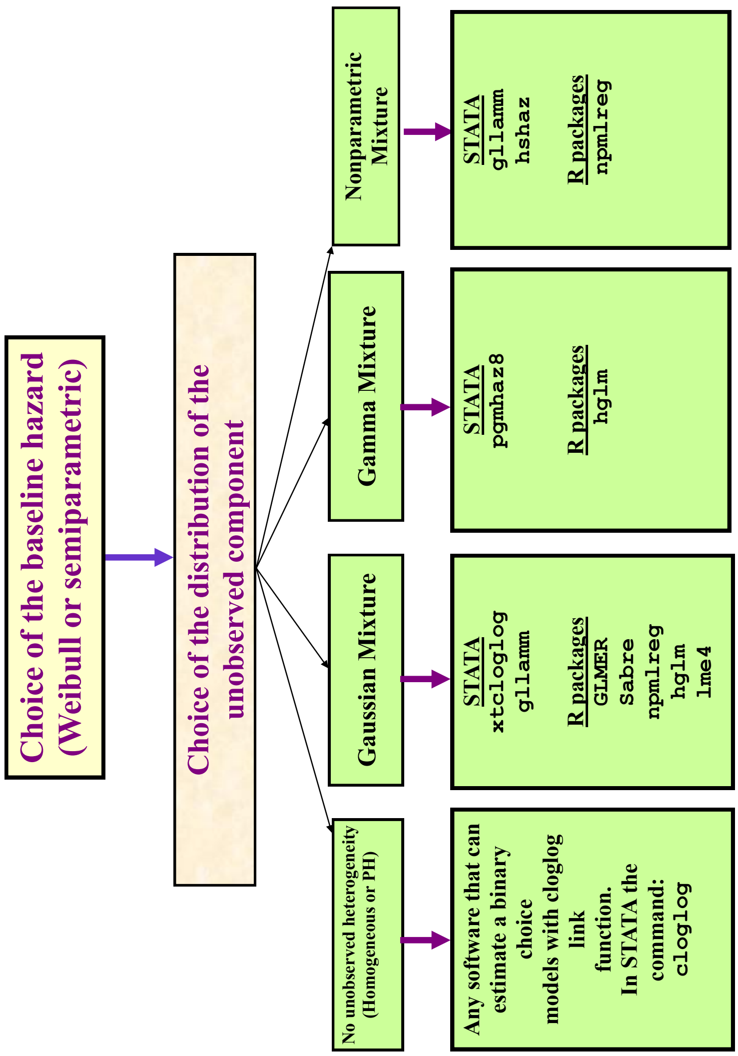


Figure B.1: Software packages to estimate (M)PH parametric models in Chapter 4.

Appendix C

Appendices of Chapter 3

C.1 MNL unrestricted models

Denote to the alternative specific coefficients of the dummy regressors by $\delta_{k(\star)}^{(\cdot)}$ and to the dummy regressors by $d_{(\star)i}^{(\cdot)}$, where $(\cdot)$ denotes to independent variable(s) and $(\star)$ denotes the value(s) of the independent variable(s) that use to generate the dummy regressor. The parametric models are specified as the following:

Linear form in age:

$$\begin{aligned} p(Y_i^d = k) &= F(\beta_{k0} + \beta_{k1} X_{age,i}^c \\ &\quad + \sum_{l=1}^{3} \delta_{kl}^{nc} d_{li}^{nc} + \delta_k^{hoh} d_i^{hoh} + \delta_k^{u5c} d_i^{u5c}). \end{aligned} \tag{C.1}$$

The quadratic form in age (Model1):

$$\begin{aligned} p(Y_i^d = k) &= F(\beta_{k0} + \beta_{k1} X_{age,i}^c + \beta_{k2} (X_{age,i}^c)^2 \\ &\quad + \sum_{l=1}^{3} \delta_{kl}^{nc} d_{li}^{nc} + \delta_k^{hoh} d_i^{hoh} + \delta_k^{u5c} d_i^{u5c}). \end{aligned} \tag{C.2}$$

Age dummies form (Model2):

$$\begin{aligned} p(Y_i^d = k) &= F(\beta_{k0} + \sum_{j=17}^{59} \delta_j^{age} d_{ji}^{age} \\ &\quad + \sum_{l=1}^{3} \delta_{kl}^{nc} d_{li}^{nc} + \delta_k^{hoh} d_i^{hoh} + \delta_k^{u5c} d_i^{u5c}). \end{aligned} \tag{C.3}$$

Age dummies with interactions 1 (Model3):

$$\begin{aligned} p(Y_i^d = k) &= F(\beta_{k0} + \sum_{j=17}^{59} \delta_j^{age} d_{ji}^{age} \\ &+ \sum_{l=1}^{3} \delta_{kl}^{nc} d_{li}^{nc} + \sum_{j=17}^{59}\sum_{l=1}^{3} \delta_{kjl}^{age,nc} d_{ijl}^{age,nc} + \delta_k^{hoh} d_i^{hoh} + \delta_k^{u5c} d_i^{u5c}). \end{aligned} \tag{C.4}$$

Age dummies with interactions 2 (Model4):

$$\begin{aligned} p(Y_i^d = k) &= F(\beta_{k0} + \sum_{j=17}^{59} \delta_j^{age} d_{ji}^{age} \\ &+ \sum_{l=1}^{3} \delta_{kl}^{nc} d_{li}^{nc} + \sum_{j=17}^{59}\sum_{l=1}^{3} \delta_{kjl}^{age,nc} d_{ijl}^{age,nc} \\ &+ \sum_{j=17}^{59} \delta_{kj}^{age,hoh} d_{ji}^{age,hoh} + \delta_k^{u5c} d_i^{u5c}). \end{aligned} \tag{C.5}$$

The models that estimated in the sample of all females add the term $\sum_{e=2}^{3} \delta_{ke}^{ed} d_{ei}^{ed}$ in the linear model, Model1 and Model2, and the term $\sum_{j=17}^{59}\sum_{e=2}^{3} \delta_{kje}^{age,ed} d_{jei}^{age,ed} + \sum_{e=2}^{3}\sum_{l=1}^{3} \delta_{kel}^{ed,nc} d_{eli}^{ed,nc}$ in Model3 and Model4.

C.2 BNL unrestricted models

Denote to the alternative specific coefficients of the dummy regressors by $\delta_{(\star)}^{(\cdot)}$ and to the dummy regressors by $d_{(\star)i}^{(\cdot)}$, where $(\cdot)$ denotes to independent variable(s) and $(\star)$ denotes the value(s) of the independent variable(s) that use to generate the dummy regressor. The parametric models are specified as the following:

Linear form in age:

$$\begin{aligned} p(U_i^d = 1) &= F(\beta_0 + \beta_1 X_{age,i}^c + \sum_{l=1}^{3} \delta_l^{nc} d_{li}^{nc} \\ &+ \delta^{hoh} d_i^{hoh} + \delta^{u5c} d_i^{u5c}). \end{aligned} \tag{C.6}$$

The quadratic form in age (Model1):

$$p(U_i^d = 1) = F(\beta_0 + \beta_1 X_{age,i}^c + \beta_2 (X_{age,i}^c)^2 + \sum_{l=1}^{3} \delta_l^{nc} d_{li}^{nc} + \delta^{hoh} d_i^{hoh} + \delta^{u5c} d_i^{u5c}). \tag{C.7}$$

Age dummies form (Model2):

$$p(U_i^d = 1) = F(\beta_0 + \sum_{j=17}^{59} \delta_j^{age} d_{ji}^{age} + \sum_{l=1}^{3} \delta_l^{nc} d_{li}^{nc} + \delta^{hoh} d_i^{hoh} + \delta^{u5c} d_i^{u5c}). \tag{C.8}$$

Age dummies with interactions 1 (Model3):

$$p(U_i^d = 1) = F(\beta_0 + \sum_{j=17}^{59} \delta_j^{age} d_{ji}^{age} + \sum_{l=1}^{3} \delta_l^{nc} d_{li}^{nc} + \sum_{j=17}^{59} \sum_{l=1}^{3} \delta_{jl}^{age,nc} d_{jli}^{age,nc} + \delta^{hoh} d_i^{hoh} + \delta^{u5c} d_i^{u5c}). \tag{C.9}$$

Age dummies with interactions 2 (Model4):

$$p(U_i^d = 1) = F(\beta_0 + \sum_{j=17}^{59} \delta_j^{age} d_{ji}^{age} + \sum_{l=1}^{3} \delta_l^{nc} d_{li}^{nc} + \sum_{j=17}^{59} \sum_{l=1}^{3} \delta_{jl}^{age,nc} d_{jli}^{age,nc} + \sum_{j=17}^{59} \delta_j^{age,hoh} d_{ji}^{age,hoh} + \delta^{u5c} d_i^{u5c}). \tag{C.10}$$

The models that estimated in the sample of all females add the term $\sum_{e=2}^{3} \delta_e^{ed} d_{ei}^{ed}$ in the linear model, Model1 and Model2, and the term $\sum_{j=17}^{59} \sum_{e=2}^{3} \delta_{je}^{age,ed} d_{jei}^{age,ed} + \sum_{e=2}^{3} \sum_{l=1}^{3} \delta_{el}^{ed,nc} d_{eli}^{ed,nc}$ in Model3 and Model4.

C.3 Replication of Zheng (2008) DGPs

In this section we attempt to reproduce some of the Monte Carlo simulations results that presented in Fan et al. (2006) paper using R programming Language. The objective of this replication is to examine the codes that are used in Chapter 2 this research to compute the smoothing specification tests of the discrete choice models are correct. However, due to the hard computations that involved in the computation of the test bootstrap distribution, the work replicates only some of the Monte Carlo simulation results that presented in Zheng (2008) paper. However, both tests are examined in our work, this doubles the computation and make the work more difficult.

The simulated model in Zheng (2008) is a Multinomial Logit that simulates a multinomial dependent variable with 3 outcomes and a continuous covariate variable that follow a standard normal distribution.

C.3.1 Replication of the Monte Carlo Simulation in Zheng (2008)

This is a replication of the results on Table 1 in Zheng (2008), but with less number of Monte Carlo replication for sample sizes 50 and 100.

Table C.1: Replication of the simulation of the size of the test in Zheng (2008)

% of H0 rejections.				
Zheng (2008)Test:-				
Sample	Monte Carlo	level of significance		
	replications	1%	5%	10%
$n = 50$	1000	1.3	4.4	9.5
$n = 100$	1000	1.2	6.4	10.6
$n = 200$	500	2	6.4	11.8
$n = 300$	500	1.6	7.4	12.2
Fan et al. (2006) Test:-				
Sample	Monte Carlo	level of significance		
	replications	1%	5%	10%
$n = 50$	1000	1.4	4.4	9.7
$n = 100$	1000	1.2	6.2	10.6
$n = 200$	500	2.2	7	11.4
$n = 300$	500	1.6	7.8	12.2

[1] The number of Monte Carlo samples is reduced than the number in the paper to reduce the computation time.

C.4 Females Labour Supply Sub-samples Models

Table C.2: Marginal effects for the MNL unrestricted models in the females with high qualifications sub-sample.

	females with high qualifications		
	Employed	Unemployed	Inactive
Panel A: Quadratic form in X^c_{age}			
$p\left(y^d \mid \overline{\mathbf{x}}\right)$	0.874	0.020	0.105
Age($= X^c_{age}$)	3.929 (0.263)	-0.373 (0.109)	-3.556 (0.241)
Age Squared	-5.153 (0.324)	0.395 (0.138)	4.758 (0.296)
Number of Children X^d_{nc}:			
1 child	-0.020 (0.011)	0.009 (0.005)	0.011 (0.010)
2 children	-0.078 (0.014)	0.005 (0.005)	0.073 (0.013)
3 or more	-0.176 (0.024)	0.010 (0.009)	0.166 (0.023)
X^d_{u5c}	-0.018 (0.008)	0.005 (0.003)	0.013 (0.008)
X^d_{hoh}	-0.150 (0.015)	-0.009 (0.003)	0.158 (0.0150
Log likelihood	-4520.30		
Pseudo R^2	0.062		
Wald stat.	553.67		
Panel B: Year dummies of X^c_{age}			
$p\left(y^d \mid \overline{\mathbf{x}}\right)$	0.877	0.018	0.105
Age dummies	the marginal effects of the age dummies are not shown		
Number of Children X^d_{nc}:			
1 child	-0.018 (0.012)	0.005 (0.004)	0.013 (0.011)
2 children	-0.068 (0.014)	0.001 (0.005)	0.067 (0.014)
3 or more	0.159 (0.024)	0.005 (0.007)	0.155 (0.024)
X^d_{u5c}	-0.017 (0.008)	0.005 (0.003)	0.013 (0.008)
X^d_{hoh}	-0.139 (0.017)	-0.005 (0.004)	0.144 (0.017)
Log likelihood	-4474.58		
Pseudo R^2	0.0715		
Wald stat.	NA		
n	10125		

[1] The basegroups are: for the Number of Children, the no children group; for the Education level, the Low Qualification group; for X^d_{u5c}, no under 5 children; X^d_{hoh}, not head of the households

[2] The MNL model is estimated for the sub-sample of the females in the working age 16-59 and the High qualification group, the predictions of this model can be compared with the predictions of the model in Table 2.6 when the dummy variable *high qual.* is one.

Table C.3: Marginal effects for the MNL unrestricted models in the females with medium qualifications sub-sample.

		females with medium qualifications	
	Employed	Unemployed	Inactive
Panel A: Quadratic form in X^c_{age}			
$p\left(y^d \left\| \overline{\mathbf{x}}\right.\right)$	0.732	0.035	0.232
Age($= X^c_{age}$)	5.553 (0.194)	-0.436 (0.076)	-5.117 (0.181)
Age Squared	-7.196 (0.270)	0.415 (0.109)	6.781 (0.253)
Number of Children X^d_{nc}:			
1 child	-0.081 (0.011)	0.006 (0.004)	0.075 (0.010)
2 children	-0.143 (0.012)	0.002 (0.004)	0.141 (0.012)
3 or more	-0.306 (0.017)	0.011 (0.006)	0.295 (0.017)
X^d_{u5c}	-0.134 (0.009)	0.028 (0.004)	0.107 (0.009)
X^d_{hoh}	-0.169 (0.012)	-0.005 (0.004)	0.175 (0.012)
Log likelihood	-10670.49		
Pseudo R^2	0.081		
Wald stat.	1609.19		
Panel B: Year dummies of X^c_{age}			
$p\left(y^d \left\| \overline{\mathbf{x}}\right.\right)$	0.735	0.034	0.231
Age dummies		the marginal effects of the age dummies are not shown	
Number of Children X^d_{nc}:			
1 child	-0.074 (0.011)	0.002 (0.004)	0.072 (0.011)
2 children	-0.134 (0.013)	-0.002 (0.004)	0.136 (0.013)
3 or more	-0.290 (0.018)	0.006 (0.006)	0.285 (0.018)
X^d_{u5c}	-0.135 (0.009)	0.028 (0.004)	0.108 (0.009)
X^d_{hoh}	-0.189 (0.014)	-0.001 (0.004)	0.190 (0.014)
Log likelihood	-10601.91		
Pseudo R^2	0.0869		
Wald stat.	1730.66		
n	16181		

[1] The basegroups are: for the Number of Children, the no children group; for the Education level, the Low Qualification group; for X^d_{u5c}, no under 5 children; X^d_{hoh}, not head of the household

[2] The MNL model is estimated for the sub-sample of the females in the working age 16-59 and the Mid qualification group, the predictions of this model can be compared with the predictions of the model in Table 2.6 when the dummy variable *ges/gcse* is one.

Table C.4: Marginal effects for the MNL unrestricted models in the females with low qualifications sub-sample.

	females with low qualifications		
	Employed	Unemployed	Inactive
Panel A: Quadratic form in X^c_{age}			
$p\left(y^d \,\vert \overline{\mathbf{x}}\right)$	0.505	0.046	0.449
Age($= X^c_{age}$)	5.092 (0.306)	-0.274 (0.103)	-4.818 (0.301)
Age Squared	-6.382 (0.397)	0.036 (0.140)	6.347 (0.393)
Number of Children X^d_{nc}:			
1 child	-0.081 (0.016)	0.003 (0.006)	0.078 (0.016)
2 children	-0.152 (0.017)	-0.011 (0.005)	0.163 (0.017)
3 or more	-0.319 (0.017)	-0.007 (0.007)	0.326 (0.018)
X^d_{u5c}	-0.116 (0.012)	0.020 (0.005)	0.096 (0.012)
X^d_{hoh}	-0.179 (0.017)	-0.017 (0.005)	0.196 (0.017)
Log likelihood	-7197.77		
Pseudo R^2	0.0722		
Wald stat.	1022.8		
Panel B: Year dummies of X^c_{age}			
$p\left(y^d \,\vert \overline{\mathbf{x}}\right)$	0.505	0.043	0.453
Age dummies		the marginal effects of the age dummies are not shown	
Number of Children X^d_{nc}:			
1 child	-0.056 (0.017)	-0.002 (0.005)	0.058 (0.017)
2 children	-0.124 (0.019)	-0.013 (0.005)	0.137 (0.019)
3 or more	-0.297 (0.019)	-0.010 (0.006)	0.307 (0.020)
X^d_{u5c}	-0.123 (0.012)	0.021 (0.005)	0.102 (0.012)
X^d_{hoh}	-0.213 (0.018)	-0.012 (0.005)	0.225 (0.018)
Log likelihood	-7132.32		
Pseudo R^2	0.0806		
Wald stat.	1074.78		
n	9005		

[1] The basegroups are: for the Number of Children, the no children group; for the Education level, the Low Qualification group; for X^d_{u5c}, no under 5 children; X^d_{hoh}, not head of the household

[2] The MNL model is estimated for the sub-sample of the females in the working age 16-59 and the Low qualification group, the predictions of this model can be compared with the predictions of the model in Table 2.6 at the basegroup.

Table C.5: Marginal effects for the probability of unemployment from the BNL models with quadratic form and with the age dummies for the females in labour force.

		sub-samples	
	High Qual.	GCE/GCSE	Low Qual
Panel A: Quadratic form in X^c_{age}			
$p\left(y^d \mid \overline{\mathbf{x}}\right)$	0.023	0.041	0.073
Age($= X^c_{age}$)	-0.506 (0.121)	-0.785 (0.088)	-1.212 (0.177)
Age Squared	0.562 (0.154)	0.839 (0.127)	1.074 (0.238)
Number of Children X^d_{nc}:			
1 child	0.011 (0.006)	0.012 (0.005)	0.023 (0.010)
2 children	0.008 (0.006)	0.012 (0.005)	0.018 (0.012)
3 or more	0.018 (0.011)	0.039 (0.011)	0.074 (0.022)
X^d_{u5c}	0.006 (0.004)	0.046 (0.006)	0.055 (0.009)
X^d_{hoh}	-0.006 (0.004)	0.002 (0.005)	-0.0001 (0.011)
Log likelihood	-1003.74	-2313.52	-1402.46
Pseudo R^2	0.019	0.0688	0.1107
Wald stat.	42.48	321.04	331.18
Panel B:Year dummies of X^c_{age}			
$p\left(y^d \mid \overline{\mathbf{x}}\right)$	0.021	0.039	0.069
Age dummies	the marginal effects of the age dummies are not shown		
Number of Children X^d_{nc}:			
1 child	0.006 (0.005)	0.008 (0.005)	0.004 (0.010)
2 children	0.003 (0.006)	0.007 (0.006)	-0.001 (0.011)
3 or more	0.010 (0.009)	0.028 (0.011)	0.041 (0.020)
X^d_{u5c}	0.006 (0.004)	0.046 (0.006)	0.059 (0.010)
X^d_{hoh}	-0.003 (0.005)	0.008 (0.006)	0.018 (0.014)
Log likelihood	-983.42	-2285.87	-1367.22
Pseudo R^2	0.0388	0.0799	0.133
Wald stat.	86.55	383.79	385.03
n	8869	12111	5027

[1] The basegroups are: for the Number of Children, the no children group; for the Education level, the Low Qualification group; for X^d_{u5c}, no under 5 children; X^d_{hoh}, not head of the household.

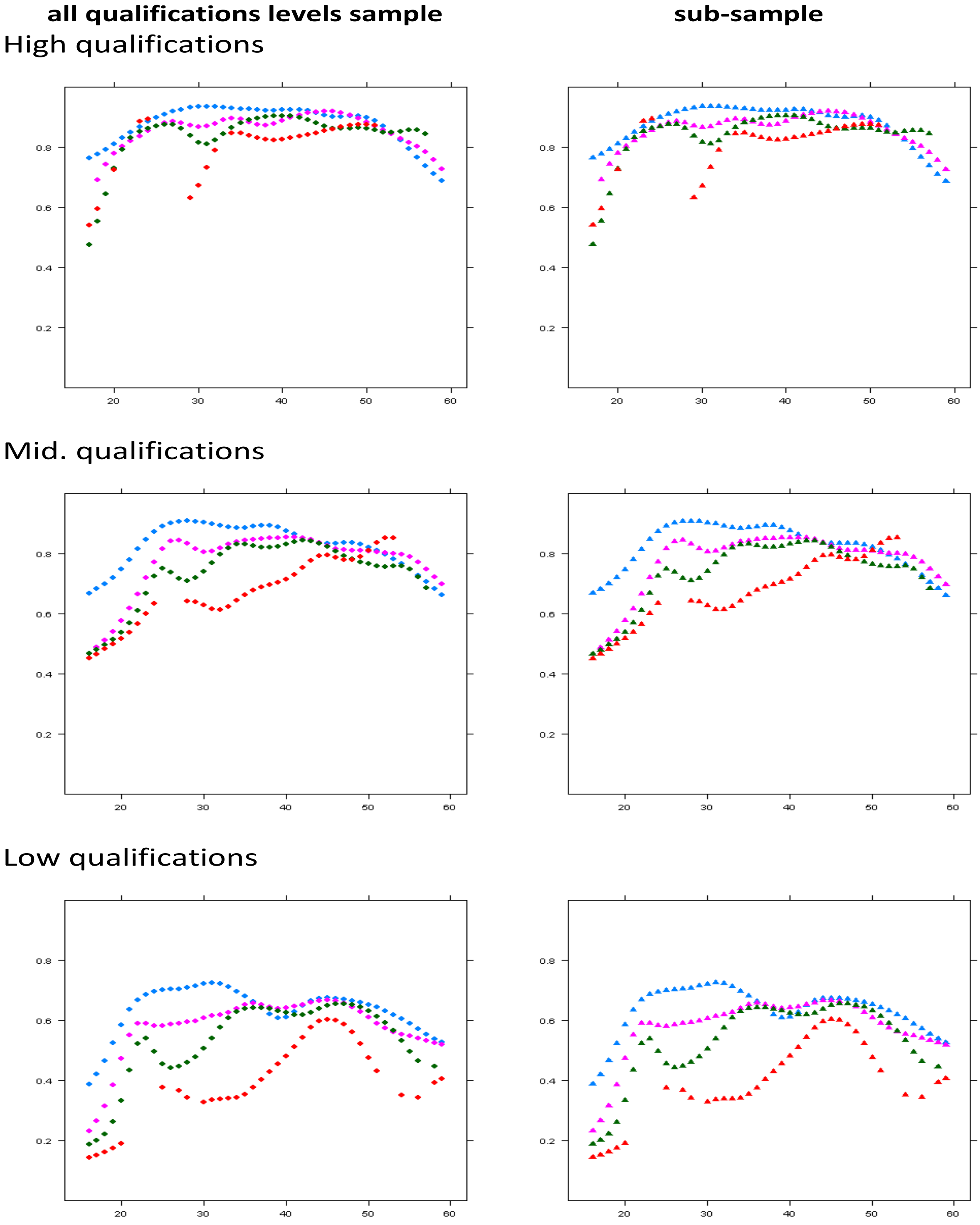


Figure C.1: Predicted probability of employment state by kernel estimator for the females with high qualifications, no children and not head of household in the sample of all females (left) compared with the predicted probability of the kernel model from the sub-sample of high qualifications (right).

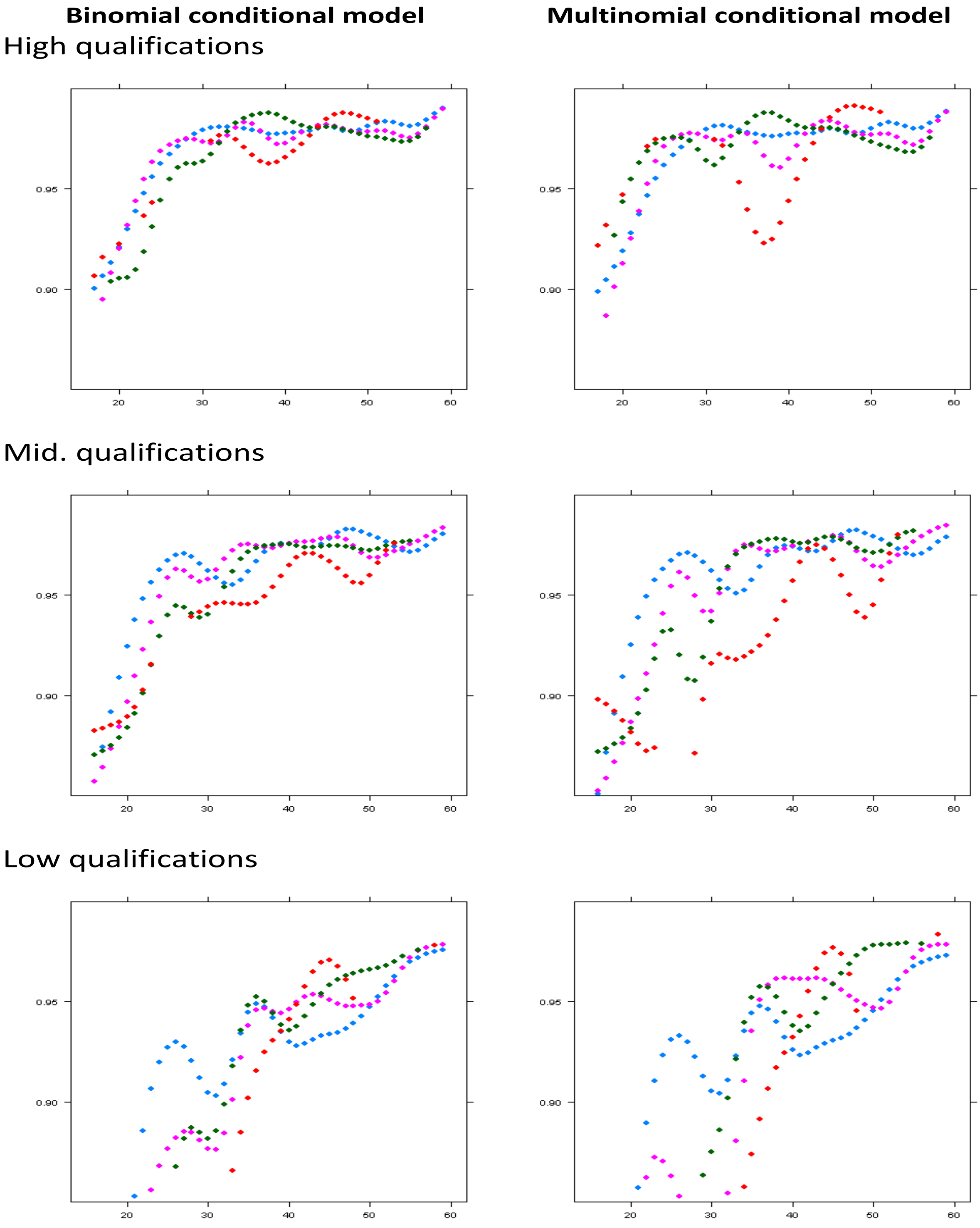


Figure C.2: Predicted probability of the employment state for the females with high qualifications, no children and not head of household as estimated in the labour for sample (Binomial conditional density - left) and in the sample of females in the working age 16-59 as a multinomial conditional density function by taking the product of the predicted conditional probabilities of employment and economically active states (right).

Appendix D

Appendices of Chapter 4

D.1 The Mathematical Proofs

Let y^c denotes the transition time variable in continuous duration model. y^c is a non-negative continuous random variable with a density f_{Y^c} and a distribution function F_{Y^c}. Let assume that Y_i^c is drawn from Weibull density function with a scale parameter λ and a shape parameter k, then the functions of Y_i^c are

$$f_{Y^c}\left(y^c;\lambda,k\right) = k\left(\frac{1}{\lambda}\right)^k \left(y^c\right)^{k-1}\exp\left[-\left(\frac{y^c}{\lambda}\right)^k\right]. \tag{D.1}$$

The distribution function is:

$$F_{Y^c}\left(y^c;\lambda,k\right) = 1-\exp\left[-\left(\frac{y^c}{\lambda}\right)^k\right]. \tag{D.2}$$

The survival function is:

$$S_{Y^c}\left(y^c\right) = \exp\left[-\left(\frac{y^c}{\lambda}\right)^k\right], \tag{D.3}$$

$$= \exp\left[-\left(\frac{1}{\lambda}\right)^k \left(y^c\right)^k\right], \tag{D.4}$$

and the hazard is

$$\phi_{Y^c}\left(y\right) = k\left(\frac{1}{\lambda}\right)^k \left(y^c\right)^{k-1}, \tag{D.5}$$

and the cumulative hazard function is

$$\Lambda_{Y^c}(y^c) = -\log\left\{\exp\left[-\left(\frac{y^c}{\lambda}\right)^k\right]\right\}, \tag{D.6}$$

$$= \left(\frac{1}{\lambda}\right)^k (y^c)^k. \tag{D.7}$$

The mean and the variance are

$$\mu_{Y^c} = \lambda\Gamma\left(1+\frac{1}{k}\right), \tag{D.8}$$

and

$$\sigma^2_{Y^c} = \lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right]. \tag{D.9}$$

Now assume that the continuous variable y^c has been grouped into intervals with equal lengths b, where b is determined exogenously, the grouped time variable j^o takes the values $1, 2, 3, ..., B$, as follows:

y^c			j^o	$f(j^o)$
$0 < y^c \leq b$	$\Rightarrow 0 < y^c \leq y^c_1$	$\Longrightarrow$	$j^o = 1$	$f(J^o = 1)$
$b < y^c \leq 2b$	$\Rightarrow y^c_1 < y^c \leq y^c_2$	$\Longrightarrow$	$j^o = 2$	$f(J^o = 2)$
$2b < y^c \leq 3b$	$\Rightarrow y^c_2 < y^c \leq y^c_3$	$\Longrightarrow$	$j^o = 3$	$f(J^o = 3)$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
$(B-1)\,b < y^c \leq Bb$	$\Rightarrow y^c_{B-1} < y^c \leq y^c_B$	$\Longrightarrow$	$j^o = B$	$f(J^o = B)$
$Bb < y^c$	$\Rightarrow y^c_B < y^c < \infty$	$\Longrightarrow$	$j^o = B+1$	$f(J^o = B+1)$

The probability mass function of the grouped time can be constructed directly from the Weibull density function as follows:

$$f(j^o) = \int_{y^c_{j^o-1}}^{y^c_{j^o}} k(\frac{1}{\lambda})u^{k-1}exp(-\left(\frac{u}{\lambda}\right)^k)du \tag{D.10}$$

Since the distribution on the Weibull variable takes the form above, the probability mass function in the first interval is

$$f(J^o = 1) = 1 - F_{Y^c}\left(y_1^c\right), \tag{D.11}$$

$$= 1 - \exp\left[-\left(\frac{y_1^c}{\lambda}\right)^k\right], \tag{D.12}$$

$$\tag{D.13}$$

and for the other intervals

$$f(J^o = j^o) = F_{Y^c}\left(y_{j^o-1}^c\right) - F_{Y^c}\left(y_{j^o}^c\right), \tag{D.14}$$

$$= S_{Y^c}\left(y_{j^o}^c\right) - S_{Y^c}\left(y_{j^o-1}^c\right), \tag{D.15}$$

$$= \exp\left[-\left(\frac{b}{\lambda}\right)^k (j^o - 1)^k\right] - \exp\left[-\left(\frac{b}{\lambda}\right)^k (j^o)^k\right]. \tag{D.16}$$

At the last interval the probability is

$$f\left(J^o = B + 1\right) = \exp\left[-\left(\frac{b}{\lambda}\right)^k B^k\right],$$

so that we can have $\sum_{j^o=1}^{B+1} f\left(j^o\right) = 1$. **This is clearly not a Weibull density function and depends crucially on the interval length** b.

Jenkins (2005) show that the hazard of the grouped time j^o at any episode $j^o \in [1, 2, ..., B]$ is:

$$\theta\left(j^o\right) = 1 - \frac{S_{Y^c}\left(j^o b\right)}{S_{Y^o}\left(\left(j^o - 1\right) b\right)}, \tag{D.17}$$

$$= 1 - \frac{\exp\left[-\left(\frac{1}{\lambda}\right)^k \left(j^o b\right)^k\right]}{\exp\left[-\left(\frac{1}{\lambda}\right)^k \left(\left(j^o - 1\right) b\right)^k\right]},$$

$$= 1 - \exp\left\{-\left(\frac{b}{\lambda}\right)^k \left[(j^o)^k - (j^o - 1)^k\right]\right\} \tag{D.18}$$

But $\Lambda_{Y^c}\left(y_{j^o}^c\right) = \left(\frac{b}{\lambda}\right)^k (j^o)^k$ and $\Lambda_{Y^c}\left(y_{j^o-1}^c\right) = \left(\frac{b}{\lambda}\right)^k (j^o - 1)^k$, then

$$\theta\left(j^o\right) = 1 - \exp\left[\Lambda_{Y^c}\left(y_{j^o-1}^c\right) - \Lambda_{Y^c}\left(y_{j^o}^c\right)\right]. \tag{D.19}$$

Then

$$1 - \theta\left(j^o\right) = \exp\left[\Lambda_{Y^c}\left(y_{j^o-1}^c\right) - \Lambda_{Y^c}\left(y_{j^o}^c\right)\right], \tag{D.20}$$

and

$$\log\left(1-\theta\left(j^{o}\right)\right)=\left[\Lambda_{Y^{c}}\left(y^{c}_{j^{o}-1}\right)-\Lambda_{Y^{c}}\left(y^{c}_{j^{o}}\right)\right], \tag{D.21}$$

$$\begin{aligned}
\log\left[-\log\left(1-\theta\left(j^{o}\right)\right)\right] &= \log\left[\left(\Lambda_{Y^{c}}\left(y^{c}_{j^{o}}\right)-\Lambda_{Y^{c}}\left(y^{c}_{j^{o}-1}\right)\right)\right], && \text{(D.22)}\\
&= \log\left[\left(\frac{b}{\lambda}\right)^{k}\left((j^{o})^{k}-(j^{o}-1)^{k}\right)\right], && \text{(D.23)}\\
&= \log\left(\frac{b}{\lambda}\right)^{k}+\log\left((j^{o})^{k}-(j^{o}-1)^{k}\right). && \text{(D.24)}
\end{aligned}$$

Which is the complementary-log-log transformation.

The proportional hazard conditional on the independent variable $\mathbf{x}$ for the baseline hazard in Eq (D.5) is given by

$$\begin{aligned}
\theta_{Y^{c}}\left(y^{c}\,|\mathbf{x}\right) &= \theta_{Y^{c}}\left(y^{c}\right)\exp(\mathbf{x}^{T}\boldsymbol{\beta}), && \text{(D.25)}\\
&= k\left(\frac{1}{\lambda}\right)^{k}(y^{c})^{k-1}\exp(\mathbf{x}^{T}\boldsymbol{\beta}), && \text{(D.26)}\\
&= k\left(\frac{1}{\lambda}\right)^{k}\left(\exp\left(\mathbf{x}^{T}\boldsymbol{\beta}\right)^{-\frac{1}{k}}\right)^{-k}(y^{c})^{k-1}, && \text{(D.27)}\\
&= k\left(\frac{1}{\lambda}\right)^{k}\left(\frac{1}{\exp\left(-k^{-1}\mathbf{x}^{T}\boldsymbol{\beta}\right)}\right)^{k}(y^{c})^{k-1}, && \text{(D.28)}\\
&= k\left(\frac{1}{\lambda\exp\left(-k^{-1}\mathbf{x}^{T}\boldsymbol{\beta}\right)}\right)^{k}(y^{c})^{k-1}, && \text{(D.29)}\\
& && \text{(D.30)}
\end{aligned}$$

Take the integration to find the cumulative hazard at the end of episode $y^{c}_{j^{o}}$.

$$\begin{aligned}
\Lambda_{Y^{c}}\left(y^{c}\,|\mathbf{x}\right) &= \int_{0}^{y^{c}_{j^{o}}} k\left(\frac{1}{\lambda\exp\left(-k^{-1}\mathbf{x}^{T}\boldsymbol{\beta}\right)}\right)^{k}(u)^{k-1}du. && \text{(D.31)}\\
&= \left(\frac{1}{\lambda\exp\left(-k^{-1}\mathbf{x}^{T}\boldsymbol{\beta}\right)}\right)^{k}(y^{c}_{j^{o}})^{k}, && \text{(D.32)}\\
&= \left(\frac{1}{\lambda\exp\left(-k^{-1}\mathbf{x}^{T}\boldsymbol{\beta}\right)}\right)^{k}(j^{o}b)^{k}. && \text{(D.33)}
\end{aligned}$$

Which is the cumulative hazard for a Weibull variable with $\left(\frac{1}{\lambda}\exp(\mathbf{x}^{T}\boldsymbol{\beta}^{*})\right)$ scale parameter.

The complementary-log-log transformation of the conditional hazard produces

$$\begin{aligned}
\log\left[-\log\left(1-\theta\left(j^{o}\left|\mathbf{x}\right.\right)\right)\right] &= \log\left[\left(\Lambda_{Y^c}\left(y^c_{j^o}\left|\mathbf{x}\right.\right)-\Lambda_{Y^c}\left(y^c_{j^o-1}\left|\mathbf{x}\right.\right)\right)\right], && \text{(D.34)}\\
&= \log\left[\left(\frac{b}{\lambda\exp\left(-k^{-1}\mathbf{x}^T\boldsymbol{\beta}\right)}\right)^k\left((j^o)^k-(j^o-1)^k\right)\right], && \text{(D.35)}\\
&= \log\left(\frac{b}{\lambda}\right)^k+\log\left((j^o)^k-(j^o-1)^k\right)+\mathbf{x}^T\boldsymbol{\beta}. && \text{(D.36)}
\end{aligned}$$

Note that $-k^{-1}\boldsymbol{\beta}$ are the coefficients of the Accelerated Failure Time regression model. The table below shows the function for the continuous Weibull transition time and the functions of the grouped time model for a fixed interval length $b = 1$.

Table D.1: Grouped time hazard and survival functions as functions on the continuous time index points

Function	Continuous time formula		Grouped time formula	
Unconditional hazard functions:				
Density	$f_{Y^c}\left(y^c_{j^o}\right) = k(\frac{1}{\lambda})(y^c_{j^o})^{k-1}exp(-\left(\frac{y^c_{j^o}}{\lambda}\right)^k)$	(1)	$f\left(j^o\right) = \int\limits_{y_{j^o-1}}^{y_{j^o}} k(\frac{1}{\lambda})u^{k-1}exp(-\left(\frac{u}{\lambda}\right)^k)du$	(2)
Distribution	$F_{Y^c}\left(y^c_{j^o}\right) = 1 - exp(-\left(\frac{y^c_{j^o}}{\lambda}\right)^k)$	(3)	$F\left(j^o\right) = \sum\limits_{r=1}^{j^o} f\left(r\right)$	(4)
Survival	$S_{Y^c}\left(y^c_{j^o}\right) = exp(-\left(\frac{y^c_{j^o}}{\lambda}\right)^k)$	(5)	$S\left(j^o\right) = 1 - \sum\limits_{r=1}^{j^o} f\left(r\right)$	(6)
Hazard	$\phi_{Y^c}\left(y^c_{j^o}\right) = k(\frac{1}{\lambda})(y^c_{j^o})^{k-1}$	(7)	$\theta\left(j^o\right) = 1 - \frac{S_Y\left(y_{j^o}\right)}{S_Y\left(y_{j^o-1}\right)}$	(8a)
			$\theta\left(j^o\right) = 1 - exp\left[-\left(\frac{1}{\lambda}\right)^k \left({j^o}^k - (j^o-1)^{k-1}\right)\right]$	(8b)
Conditional hazard functions:				
Density	$f_{Y^c\|\mathbf{x}}\left(y^c_{j^o}\|\mathbf{x}\right) = exp\left(\mathbf{x}^T\boldsymbol{\beta}\right) f_Y\left(y^c_{j^o}\right)^{exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}$	(9)	$f\left(j^o\|\mathbf{x}\right) = F_Y\left(y_{j^o}\right)^{exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)} - F_Y\left(y_{j^o-1}\right)^{exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}$	(10)
Distribution	$F_{Y^c\|\mathbf{x}}\left(y^c_{j^o}\|\mathbf{x}\right) = F_Y\left(y^c_{j^o}\right)^{exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}$	(11)	$F\left(j^o\|\mathbf{x}\right) = F_Y\left(y_{j^o}\right)^{exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}$	(12)
Survival	$S_{Y^c\|\mathbf{x}}\left(y^c_{j^o}\|\mathbf{x}\right) = 1 - F_Y\left(y^c_{j^o}\right)^{exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}$	(13)	$S\left(j^o\|\mathbf{x}\right) = S_Y\left(y_{j^o}\right)^{exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}$	(14)
Hazard	$\phi_{Y^c}\left(y^c_{j^o}\|\mathbf{x}\right) = h_Y\left(y^c_{j^o}\right) exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)$	(15)	$\theta\left(j^o\|\mathbf{x}\right) = 1 - exp\left[-exp\left(\mathbf{x}^T\boldsymbol{\beta} + klog\left(\frac{1}{\lambda}\right)\right) + log\left({j^o}^k - (j^o-1)^{k-1}\right)\right]$	(16)

[1] The grouped Weibull is denoted j^o, is a discrete ordered variable that measures the number of intervals that elapsed in the spell.
[2] All the formula are simplified by setting the interval length $b = 1$. The interval upper limit is $y^c_{j^o} = bj^o = j^o$.

D.2 The Simulation Results

Table D.3: aMSE for the estimation techniques for the unconditional hazard rate for Weibull with shape=1 and scale=1.

	n	$\widehat{\pi}$	aMSE($\widehat{\theta}^{WPH}(j)$)	aMSE($\widehat{\theta}^{SPH}(j)$)	aMSE($\widehat{\theta}^{EKH}(j)$)	aMSE($\widehat{\theta}^{DNK}(j)$)
1	100	0.0164%	0.005332	0.003446	0.009426	0.011955
2	200	0.0164%	0.003074	0.001853	0.005548	0.005856
3	500	0.0164%	0.001267	0.000752	0.002287	0.002262
4	1000	0.0164%	0.000646	0.000367	0.001176	0.001166
5	100	2%	0.005753	0.003476	0.009367	0.012001
6	200	2%	0.003390	0.001874	0.006208	0.007338
7	500	2%	0.001262	0.000753	0.002322	0.002330
8	1000	2%	0.000678	0.000409	0.001101	0.001143
9	100	4.999%	0.006360	0.004239	0.011767	0.012860
10	200	4.999%	0.003691	0.002390	0.007824	0.008091
11	500	4.999%	0.001650	0.001059	0.003136	0.003068
12	1000	4.999%	0.000185	0.000272	0.000655	0.000720
13	100	9.999%	0.007935	0.005008	0.015788	0.013113
14	200	9.999%	0.005100	0.003009	0.010547	0.011431
15	500	9.999%	0.002541	0.001772	0.005531	0.005780
16	1000	9.999%	0.001887	0.001412	0.003187	0.003553
17	100	24.99%	0.023973	0.016507	0.085253	0.019767
18	200	24.99%	0.018075	0.012555	0.072481	0.036933
19	500	24.99%	0.015399	0.011029	0.065739	0.053787
20	1000	24.99%	0.013685	0.009819	0.063792	0.054100

[1] The average mean square error (aMES) is computed for the models:$\widehat{\theta}^{WPH}(j)$, the grouped time hazard with $log(j)$ specification of the elapsed time in the spell variable. $\widehat{\theta}^{SPH}(j)$, the piecewise grouped time hazard. $\widehat{\theta}^{EKH}(j)$ kernel external hazard model, and $\widehat{\theta}^{DNK}(j)$ the density implied hazard.

Table D.2: Coefficients of parametric models for the unconditional hazard for Weibull with shape=1 and scale=1.

	n	$\widehat{\pi}$	$\widehat{k}^{AFT}$	$\widehat{\lambda}^{AFT}$	$\widehat{\alpha}_0^{WPH}$	$\widehat{\alpha}_{log(j)}^{WPH}$	$\widehat{\gamma}_0^{naive}$
1	100	0.0164%	1.016	1.0010	-0.002183	0.044480	0.045000
2	200	0.0164%	1.008	0.9982	-0.000137	0.022940	0.022512
3	500	0.0164%	1.003	1.0030	-0.001791	0.004742	0.009683
4	1000	0.0164%	1.002	0.9985	-0.000443	0.010050	0.004298
5	100	2%	1.014	1.0000	-0.024800	0.062850	0.040372
6	200	2%	1.009	0.9976	-0.017700	0.041060	0.021031
7	500	2%	1.003	0.9982	-0.011310	0.003260	0.008262
8	1000	2%	1.003	1.0000	-0.013440	0.002428	0.004126
9	100	4.999%	1.007	1.0030	-0.047310	0.054550	0.035995
10	200	4.999%	1.005	1.0060	-0.042040	0.010860	0.019897
11	500	4.999%	1.005	1.0030	-0.035400	-0.002882	0.007945
12	1000	4.999%	1.002	0.9984	-0.034690	0.002387	0.003947
13	100	9.999%	1.018	1.0050	-0.076040	0.028980	0.032920
14	200	9.999%	1.007	0.9964	-0.067110	-0.002553	0.017219
15	500	9.999%	1.001	1.0000	-0.071690	-0.010090	0.006229
16	1000	9.999%	1.001	1.0000	-0.071300	-0.016780	0.003018
17	100	24.99%	1.019	1.0090	-0.191400	-0.122800	0.011004
18	200	24.99%	1.010	1.0010	-0.178300	-0.149700	0.003309
19	500	24.99%	1.002	1.0040	-0.181300	-0.161900	0.000073
20	1000	24.99%	1.000	0.9975	-0.177200	-0.158000	0.000005

[1] $\widehat{\pi}$ is the proportion of the right censored observations. $\widehat{\lambda}^{AFT}$ and $\widehat{k}^{AFT}$ are the estimated scale and shape parameters from the continuous time AFT. $\widehat{\alpha}_0^{WPH}$ and $\widehat{\alpha}_{log(j)}^{WPH}$ are the coefficients of the grouped time PH model. $\widehat{\gamma}_0^{naive}$ is the MLCV naive bandwidth of grouped time.
[2] The number of the Monte Carlo samples is 500 in all simulations.
[3] aMSE for the models in this table are shown in Table D.3 below.

Table D.4: Coefficients of parametric models for the unconditional hazard for Weibull with shape=1 and scale=0.75.

	n	$\widehat{\pi}$	$\widehat{k}^{AFT}$	$\widehat{\lambda}^{AFT}$	$\widehat{\alpha}_0^{WPH}$	$\widehat{\alpha}_{log(j)}^{WPH}$	$\widehat{\gamma}_0^{naive}$
1	100	0.0191%	0.7627	1.0050	-0.02435	-0.3850	0.091761
2	200	0.0191%	0.7562	1.0010	-0.01491	-0.4145	0.057789
3	500	0.0191%	0.7533	1.0010	-0.01689	-0.4223	0.025770
4	1000	0.0191%	0.7490	0.9985	-0.01334	-0.4349	0.013252
5	100	1.984%	0.7613	1.0050	-0.03157	-0.3963	0.090620
6	200	1.984%	0.7567	1.0060	-0.03057	-0.4128	0.054975
7	500	1.984%	0.7541	1.0000	-0.02402	-0.4290	0.023246
8	1000	1.984%	0.7523	1.0010	-0.02398	-0.4337	0.013635
9	100	5.002%	0.7618	1.0050	-0.04287	-0.4071	0.083544
10	200	5.002%	0.7528	1.0040	-0.04022	-0.4357	0.047524
11	500	5.002%	0.7515	0.9995	-0.03704	-0.4406	0.020196
12	1000	5.002%	0.7520	1.0010	-0.03677	-0.4428	0.009969
13	100	9.99%	0.7614	1.0020	-0.06529	-0.4255	0.060011
14	200	9.99%	0.7540	1.0030	-0.06615	-0.4499	0.030697
15	500	9.99%	0.7479	0.9959	-0.05464	-0.4754	0.012490
16	1000	9.99%	0.7495	1.0000	-0.05707	-0.4771	0.005104
17	100	24.99%	0.7601	1.0050	-0.13370	-0.6295	0.015328
18	200	24.99%	0.7570	1.0090	-0.14000	-0.6253	0.005592
19	500	24.99%	0.7547	1.0040	-0.13820	-0.6345	0.001359
20	1000	24.99%	0.7508	0.9976	-0.13220	-0.6460	0.000087

[1] $\widehat{\pi}$ is the proportion of the right censored observations. $\widehat{\lambda}^{AFT}$ and $\widehat{k}^{AFT}$ are the estimated scale and shape parameters from the continuous time AFT. $\widehat{\alpha}_0^{WPH}$ and $\widehat{\alpha}_{log(j)}^{WPH}$ are the coefficients of the grouped time PH model. $\widehat{\gamma}_0^{naive}$ is the MLCV naive bandwidth of grouped time.

[2] The number of the Monte Carlo samples is 500 in all simulations.

[3] aMSE for the models in this table are shown in Table D.5 below.

Table D.5: aMSE for the estimation techniques for the unconditional hazard rate with shape=1 and scale=0.75.

	n	$\widehat{\pi}$	aMSE($\widehat{\theta}^{WPH}(j)$)	aMSE($\widehat{\theta}^{SPH}(j)$)	aMSE($\widehat{\theta}^{EKH}(j)$)	aMSE($\widehat{\theta}^{DNK}(j)$)
1	100	0.0191%	0.004566	0.035007	0.025352	0.027044
2	200	0.0191%	0.002531	0.018283	0.016979	0.017891
3	500	0.0191%	0.001181	0.006735	0.006717	0.006691
4	1000	0.0191%	0.000807	0.003031	0.003049	0.003026
5	100	1.984%	0.004798	0.034868	0.025591	0.027135
6	200	1.984%	0.002907	0.019885	0.017602	0.019495
7	500	1.984%	0.001184	0.007240	0.006828	0.006804
8	1000	1.984%	0.000910	0.003515	0.003518	0.003619
9	100	5.002%	0.005285	0.041025	0.029924	0.029761
10	200	5.002%	0.003035	0.022921	0.019868	0.021493
11	500	5.002%	0.001653	0.008238	0.008138	0.009535
12	1000	5.002%	0.001257	0.004008	0.004043	0.004761
13	100	9.99%	0.008072	0.050095	0.036690	0.039242
14	200	9.99%	0.004429	0.030821	0.025480	0.032618
15	500	9.99%	0.002971	0.013357	0.013408	0.020442
16	1000	9.99%	0.002618	0.006508	0.007423	0.013472
17	100	24.99%	0.017565	0.066043	0.061106	0.089800
18	200	24.99%	0.012988	0.050071	0.046062	0.097669
19	500	24.99%	0.010906	0.038705	0.038771	0.113423
20	1000	24.99%	0.010596	0.032627	0.035192	0.124522

[1] The average mean square error (aMES) is computed for the models:$\widehat{\theta}^{WPH}(j)$, the grouped time hazard with $log(j)$ specification of the elapsed time in the spell variable. $\widehat{\theta}^{SPH}(j)$, the piecewise grouped time hazard. $\widehat{\theta}^{EKH}(j)$ kernel external hazard model, and $\widehat{\theta}^{DNK}(j)$ the density implied hazard.

Table D.6: Coefficients of parametric models for the unconditional hazard for Weibull with shape=1 and scale=0.50.

	n	$\widehat{\pi}$	$\widehat{k}^{AFT}$	$\widehat{\lambda}^{AFT}$	$\widehat{\alpha}_0^{WPH}$	$\widehat{\alpha}_{log(j)}^{WPH}$	$\widehat{\gamma}_0^{naive}$
1	100	0.0275%	0.5050	1.0240	-0.06972	-0.8189	0.268743
2	200	0.0275%	0.5041	1.0110	-0.06120	-0.8277	0.192186
3	500	0.0275%	0.5012	1.0010	-0.06148	-0.8304	0.118166
4	1000	0.0275%	0.5004	1.0030	-0.06411	-0.8307	0.073156
5	100	2.002%	0.5053	1.0130	-0.06507	-0.8302	0.236345
6	200	2.002%	0.5066	1.0180	-0.06696	-0.8343	0.168660
7	500	2.002%	0.5001	1.0020	-0.06278	-0.8455	0.096857
8	1000	2.002%	0.5004	1.0020	-0.06207	-0.8461	0.058277
9	100	4.994%	0.5075	1.0200	-0.06131	-0.8572	0.193049
10	200	4.994%	0.5028	1.0060	-0.06005	-0.8673	0.133222
11	500	4.994%	0.5011	1.0040	-0.06063	-0.8723	0.062834
12	1000	4.994%	0.5008	1.0020	-0.06063	-0.8733	0.032309
13	100	9.984%	0.5067	1.0260	-0.06510	-0.9160	0.135402
14	200	9.984%	0.5056	1.0090	-0.06001	-0.9199	0.066573
15	500	9.984%	0.4995	0.9962	-0.05497	-0.9347	0.023784
16	1000	9.984%	0.4996	1.0020	-0.05996	-0.9317	0.007589
17	100	25%	0.5074	1.0190	-0.09356	-1.1700	0.017785
18	200	25%	0.5024	1.0200	-0.08899	-1.2110	0.005187
19	500	25%	0.4989	1.0060	-0.08955	-1.1890	0.000389
20	1000	25%	0.5001	0.9942	-0.08453	-1.1920	0.000011

[1] $\widehat{\pi}$ is the proportion of the right censored observations. $\widehat{\lambda}^{AFT}$ and $\widehat{k}^{AFT}$ are the estimated scale and shape parameters from the continuous time AFT. $\widehat{\alpha}_0^{WPH}$ and $\widehat{\alpha}_{log(j)}^{WPH}$ are the coefficients of the grouped time PH model. $\widehat{\gamma}_0^{naive}$ is the MLCV naive bandwidth of grouped time.
[2] The number of the Monte Carlo samples is 500 in all simulations.
[3] aMSE for the models in this table are shown in Table D.7 below.

Table D.7: aMSE for the estimation techniques for the unconditional hazard rate with shape=1 and scale=0.50.

	n	$\widehat{\pi}$	aMSE($\widehat{\theta}^{WPH}(j)$)	aMSE($\widehat{\theta}^{SPH}(j)$)	aMSE($\widehat{\theta}^{EKH}(j)$)	aMSE($\widehat{\theta}^{DNK}(j)$)
1	100	0.0275%	0.001492	0.016232	0.007148	0.007220
2	200	0.0275%	0.001189	0.007870	0.004840	0.005117
3	500	0.0275%	0.000979	0.002444	0.002058	0.002386
4	1000	0.0275%	0.000927	0.001040	0.001040	0.001225
5	100	2.002%	0.001717	0.018696	0.008146	0.007820
6	200	2.002%	0.001252	0.010083	0.005116	0.004951
7	500	2.002%	0.001102	0.003008	0.002318	0.002472
8	1000	2.002%	0.001034	0.001232	0.001255	0.001352
9	100	4.994%	0.002233	0.021147	0.008848	0.007810
10	200	4.994%	0.001653	0.011842	0.006145	0.005756
11	500	4.994%	0.001352	0.004982	0.003120	0.003336
12	1000	4.994%	0.001265	0.002152	0.001738	0.002233
13	100	9.984%	0.003135	0.023886	0.011361	0.010249
14	200	9.984%	0.002275	0.014012	0.007798	0.012591
15	500	9.984%	0.001997	0.006992	0.004825	0.018739
16	1000	9.984%	0.001874	0.003793	0.003049	0.027478
17	100	25%	0.006175	0.019693	0.016626	0.168449
18	200	25%	0.005462	0.011538	0.012449	0.179391
19	500	25%	0.004488	0.006056	0.008242	0.187374
20	1000	25%	0.004246	0.004809	0.006648	0.187418

[1] The average mean square error (aMES) is computed for the models:$\widehat{\theta}^{WPH}(j)$, the grouped time hazard with $log(j)$ specification of the elapsed time in the spell variable. $\widehat{\theta}^{SPH}(j)$, the piecewise grouped time hazard. $\widehat{\theta}^{EKH}(j)$ kernel external hazard model, and $\widehat{\theta}^{DNK}(j)$ the density implied hazard.

Table D.8: Coefficients of parametric models for the unconditional hazard for Weibull with shape=1 and scale=0.25.

	n	$\widehat{\pi}$	$\widehat{k}^{AFT}$	$\widehat{\lambda}^{AFT}$	$\widehat{\alpha}_0^{WPH}$	$\widehat{\alpha}_{log(j)}^{WPH}$	$\widehat{\gamma}_0^{naive}$
1	100	0.027%	0.2538	1.073	-0.22710	-1.098	0.629026
2	200	0.027%	0.2524	1.039	-0.22890	-1.099	0.472990
3	500	0.027%	0.2514	1.015	-0.22490	-1.102	0.292734
4	1000	0.027%	0.2505	1.004	-0.22640	-1.102	0.183169
5	100	1.991%	0.2540	1.032	-0.19490	-1.139	0.643352
6	200	1.991%	0.2522	1.042	-0.20600	-1.140	0.489129
7	500	1.991%	0.2502	1.008	-0.20250	-1.145	0.295361
8	1000	1.991%	0.2505	1.005	-0.20020	-1.145	0.184509
9	100	4.993%	0.2531	1.108	-0.17920	-1.199	0.647731
10	200	4.993%	0.2509	1.070	-0.18350	-1.201	0.514523
11	500	4.993%	0.2506	1.010	-0.16900	-1.207	0.312028
12	1000	4.993%	0.2507	1.010	-0.17240	-1.204	0.189034
13	100	9.996%	0.2513	1.061	-0.12780	-1.303	0.459381
14	200	9.996%	0.2513	1.021	-0.12900	-1.295	0.324829
15	500	9.996%	0.2502	1.016	-0.13770	-1.291	0.165767
16	1000	9.996%	0.2499	1.007	-0.13730	-1.292	0.085604
17	100	25%	0.2544	1.110	-0.07810	-1.571	0.138790
18	200	25%	0.2514	1.057	-0.07462	-1.588	0.055947
19	500	25%	0.2512	1.007	-0.06934	-1.578	0.004967
20	1000	25%	0.2502	0.993	-0.06991	-1.577	0.000178

[1] $\widehat{\pi}$ is the proportion of the right censored observations. $\widehat{\lambda}^{AFT}$ and $\widehat{k}^{AFT}$ are the estimated scale and shape parameters from the continuous time AFT. $\widehat{\alpha}_0^{WPH}$ and $\widehat{\alpha}_{log(j)}^{WPH}$ are the coefficients of the grouped time PH model. $\widehat{\gamma}_0^{naive}$ is the MLCV naive bandwidth of grouped time.
[2] The number of the Monte Carlo samples is 500 in all simulations.
[3] aMSE for the models in this table are shown in Table D.9 below.

Table D.9: aMSE for the estimation techniques for the unconditional hazard rate with shape=1 and scale=0.25.

	n	$\widehat{\pi}$	aMSE($\widehat{\theta}^{WPH}(j)$)	aMSE($\widehat{\theta}^{SPH}(j)$)	aMSE($\widehat{\theta}^{EKH}(j)$)	aMSE($\widehat{\theta}^{DNK}(j)$)
1	100	0.027%	0.001019	0.000967	0.002961	0.003001
2	200	0.027%	0.000960	0.000457	0.001664	0.001598
3	500	0.027%	0.000926	0.000183	0.000671	0.000653
4	1000	0.027%	0.000914	0.000093	0.000291	0.000307
5	100	1.991%	0.000952	0.001048	0.003098	0.003134
6	200	1.991%	0.000870	0.000471	0.001783	0.001715
7	500	1.991%	0.000829	0.000187	0.000686	0.000657
8	1000	1.991%	0.000819	0.000095	0.000293	0.000304
9	100	4.993%	0.000817	0.001024	0.003171	0.003204
10	200	4.993%	0.000747	0.000475	0.001967	0.001889
11	500	4.993%	0.000719	0.000191	0.000754	0.000711
12	1000	4.993%	0.000711	0.000098	0.000317	0.000326
13	100	9.996%	0.000703	0.001192	0.001899	0.001639
14	200	9.996%	0.000646	0.000514	0.000997	0.000913
15	500	9.996%	0.000606	0.000197	0.000356	0.000437
16	1000	9.996%	0.000593	0.000101	0.000162	0.000264
17	100	25%	0.001124	0.004904	0.003446	0.032254
18	200	25%	0.000979	0.002327	0.002038	0.054539
19	500	25%	0.000893	0.000923	0.000902	0.072925
20	1000	25%	0.000863	0.000487	0.000485	0.074671

[1] The average mean square error (aMES) is computed for the models:$\widehat{\theta}^{WPH}(j)$, the grouped time hazard with $log(j)$ specification of the elapsed time in the spell variable. $\widehat{\theta}^{SPH}(j)$, the piecewise grouped time hazard. $\widehat{\theta}^{EKH}(j)$ kernel external hazard model, and $\widehat{\theta}^{DNK}(j)$ the density implied hazard.

Table D.10: Coefficients of parametric models for the unconditional hazard for Weibull with shape=3 and scale=1.

	n	$\widehat{\pi}$	$\widehat{k}^{AFT}$	$\widehat{\lambda}^{AFT}$	$\widehat{\alpha}_0^{WPH}$	$\widehat{\alpha}_{log(j)}^{WPH}$	$\widehat{\gamma}_0^{naive}$
1	100	0.0164%	1.0170	3.009	-1.112	0.020560	0.235593
2	200	0.0164%	1.0090	3.007	-1.113	0.017770	0.151203
3	500	0.0164%	1.0020	2.994	-1.099	0.003680	0.075134
4	1000	0.0164%	0.9997	3.003	-1.099	-0.000111	0.038404
5	100	2%	1.0180	3.008	-1.125	0.032070	0.231950
6	200	2%	1.0090	2.993	-1.111	0.016020	0.151197
7	500	2%	1.0030	3.012	-1.109	0.004786	0.074217
8	1000	2%	1.0010	3.002	-1.106	0.003932	0.038204
9	100	4.999%	1.0130	2.979	-1.114	0.022590	0.225576
10	200	4.999%	1.0070	3.007	-1.117	0.010580	0.139404
11	500	4.999%	1.0030	2.992	-1.110	0.006075	0.065851
12	1000	4.999%	1.0000	3.002	-1.110	0.001558	0.035393
13	100	9.999%	1.0140	2.994	-1.126	0.019030	0.203409
14	200	9.999%	1.0070	2.998	-1.124	0.009652	0.123773
15	500	9.999%	1.0040	3.018	-1.124	0.001028	0.053759
16	1000	9.999%	1.0020	2.998	-1.118	0.000542	0.028344
17	100	24.99%	1.0220	3.021	-1.164	0.004123	0.112877
18	200	24.99%	1.0060	3.009	-1.148	-0.015710	0.051956
19	500	24.99%	0.9988	3.015	-1.143	-0.028320	0.011311
20	1000	24.99%	0.9998	3.004	-1.140	-0.026700	0.002477

[1] $\widehat{\pi}$ is the proportion of the right censored observations. $\widehat{\lambda}^{AFT}$ and $\widehat{k}^{AFT}$ are the estimated scale and shape parameters from the continuous time AFT. $\widehat{\alpha}_0^{WPH}$ and $\widehat{\alpha}_{log(j)}^{WPH}$ are the coefficients of the grouped time PH model. $\widehat{\gamma}_0^{naive}$ is the MLCV naive bandwidth of grouped time.

[2] The number of the Monte Carlo samples is 500 in all simulations.

[3] aMSE for the models in this table are shown in Table D.11 below.

Table D.11: aMSE for the estimation techniques for the unconditional hazard rate ith shape=3 and scale=1.

	n	$\widehat{\pi}$	aMSE($\widehat{\theta}^{WPH}(j)$)	aMSE($\widehat{\theta}^{SPH}(j)$)	aMSE($\widehat{\theta}^{EKH}(j)$)	aMSE($\widehat{\theta}^{DNK}(j)$)
1	100	0.0164%	0.001867	0.002405	0.002776	0.005279
2	200	0.0164%	0.000896	0.001151	0.001955	0.002960
3	500	0.0164%	0.000389	0.000504	0.001190	0.001321
4	1000	0.0164%	0.000176	0.000238	0.000552	0.000546
5	100	2%	0.002091	0.002891	0.002832	0.005381
6	200	2%	0.001035	0.001265	0.002005	0.003125
7	500	2%	0.000409	0.000530	0.001255	0.001514
8	1000	2%	0.000201	0.000261	0.000586	0.000575
9	100	4.999%	0.001959	0.002861	0.003386	0.005269
10	200	4.999%	0.001012	0.001482	0.002190	0.003288
11	500	4.999%	0.000431	0.000560	0.001299	0.001588
12	1000	4.999%	0.000185	0.000272	0.000655	0.000720
13	100	9.999%	0.002253	0.003589	0.004665	0.005101
14	200	9.999%	0.001138	0.001566	0.002658	0.002863
15	500	9.999%	0.000476	0.000669	0.001609	0.002607
16	1000	9.999%	0.000250	0.000335	0.000882	0.001909
17	100	24.99%	0.003243	0.005895	0.017501	0.005457
18	200	24.99%	0.001600	0.003026	0.014219	0.005062
19	500	24.99%	0.001008	0.001533	0.012015	0.013452
20	1000	24.99%	0.000710	0.001092	0.010526	0.019898

[1] The average mean square error (aMES) is computed for the models:$\widehat{\theta}^{WPH}(j)$, the grouped time hazard with $log(j)$ specification of the elapsed time in the spell variable. $\widehat{\theta}^{SPH}(j)$, the piecewise grouped time hazard. $\widehat{\theta}^{EKH}(j)$ kernel external hazard model, and $\widehat{\theta}^{DNK}(j)$ the density implied hazard.

Table D.12: Coefficients of parametric models for the unconditional hazard for Weibull with shape=3 and scale=0.75.

	n	$\widehat{\pi}$	$\widehat{k}^{AFT}$	$\widehat{\lambda}^{AFT}$	$\widehat{\alpha}_0^{WPH}$	$\widehat{\alpha}_{log(j)}^{WPH}$	$\widehat{\gamma}_0^{naive}$
1	100	0.0191%	0.7541	2.991	-0.8753	-0.3555	0.336117
2	200	0.0191%	0.7590	3.009	-0.8814	-0.3524	0.245971
3	500	0.0191%	0.7516	2.992	-0.8701	-0.3657	0.136030
4	1000	0.0191%	0.7506	2.992	-0.8729	-0.3642	0.079472
5	100	1.984%	0.7625	3.020	-0.8913	-0.3440	0.322385
6	200	1.984%	0.7514	3.010	-0.8789	-0.3635	0.235871
7	500	1.984%	0.7538	3.010	-0.8829	-0.3596	0.134903
8	1000	1.984%	0.7523	3.000	-0.8772	-0.3639	0.075323
9	100	5.002%	0.7569	3.029	-0.8872	-0.3621	0.307053
10	200	5.002%	0.7583	3.009	-0.8862	-0.3579	0.214290
11	500	5.002%	0.7549	3.014	-0.8845	-0.3639	0.115333
12	1000	5.002%	0.7503	3.005	-0.8784	-0.3711	0.065779
13	100	9.99%	0.7577	2.983	-0.8728	-0.3745	0.267615
14	200	9.99%	0.7555	3.000	-0.8799	-0.3756	0.173714
15	500	9.99%	0.7531	3.014	-0.8851	-0.3764	0.080780
16	1000	9.99%	0.7502	2.996	-0.8766	-0.3827	0.040039
17	100	24.99%	0.7610	3.029	-0.8947	-0.4159	0.121934
18	200	24.99%	0.7524	3.027	-0.8924	-0.4264	0.044784
19	500	24.99%	0.7506	3.005	-0.8863	-0.4298	0.006108
20	1000	24.99%	0.7507	3.017	-0.8881	-0.4313	0.001076

[1] $\widehat{\pi}$ is the proportion of the right censored observations. $\widehat{\lambda}^{AFT}$ and $\widehat{k}^{AFT}$ are the estimated scale and shape parameters from the continuous time AFT. $\widehat{\alpha}_0^{WPH}$ and $\widehat{\alpha}_{log(j)}^{WPH}$ are the coefficients of the grouped time PH model. $\widehat{\gamma}_0^{naive}$ is the MLCV naive bandwidth of grouped time.
[2] The number of the Monte Carlo samples is 500 in all simulations.
[3] aMSE for the models in this table are shown in Table D.13 below.

Table D.13: aMSE for the estimation techniques for the unconditional hazard rate with shape=3 and scale=0.75.

	n	$\widehat{\pi}$	aMSE($\widehat{\theta}^{WPH}(j)$)	aMSE($\widehat{\theta}^{SPH}(j)$)	aMSE($\widehat{\theta}^{EKH}(j)$)	aMSE($\widehat{\theta}^{DNK}(j)$)
1	100	0.0191%	0.000865	0.009819	0.004797	0.006013
2	200	0.0191%	0.000478	0.005106	0.003375	0.004317
3	500	0.0191%	0.000269	0.001434	0.001659	0.002299
4	1000	0.0191%	0.000193	0.000696	0.001039	0.001448
5	100	1.984%	0.001008	0.010697	0.005356	0.006289
6	200	1.984%	0.000496	0.005720	0.003526	0.004290
7	500	1.984%	0.000269	0.001885	0.001876	0.002412
8	1000	1.984%	0.000197	0.000807	0.001167	0.001506
9	100	5.002%	0.000995	0.011332	0.005592	0.006240
10	200	5.002%	0.000535	0.006963	0.003939	0.004308
11	500	5.002%	0.000294	0.002188	0.002369	0.002645
12	1000	5.002%	0.000240	0.000884	0.001485	0.001676
13	100	9.99%	0.001230	0.014422	0.006346	0.006309
14	200	9.99%	0.000696	0.009790	0.004952	0.005055
15	500	9.99%	0.000411	0.004723	0.003377	0.003866
16	1000	9.99%	0.000331	0.001605	0.002323	0.002722
17	100	24.99%	0.001960	0.009276	0.008316	0.012540
18	200	24.99%	0.001312	0.004445	0.005774	0.026139
19	500	24.99%	0.000846	0.001907	0.003870	0.050913
20	1000	24.99%	0.000708	0.001005	0.002980	0.064090

[1] The average mean square error (aMES) is computed for the models:$\widehat{\theta}^{WPH}(j)$, the grouped time hazard with $log(j)$ specification of the elapsed time in the spell variable. $\widehat{\theta}^{SPH}(j)$, the piecewise grouped time hazard. $\widehat{\theta}^{EKH}(j)$ kernel external hazard model, and $\widehat{\theta}^{DNK}(j)$ the density implied hazard.

Table D.14: Coefficients of parametric models for the unconditional hazard for Weibull with shape=3 and scale=0.50.

	n	$\widehat{\pi}$	$\widehat{k}^{AFT}$	$\widehat{\lambda}^{AFT}$	$\widehat{\alpha}_0^{WPH}$	$\widehat{\alpha}_{log(j)}^{WPH}$	$\widehat{\gamma}_0^{naive}$
1	100	0.0275%	0.5061	3.031	-0.6794	-0.7216	0.531526
2	200	0.0275%	0.5055	3.021	-0.6842	-0.7198	0.452056
3	500	0.0275%	0.5024	3.009	-0.6833	-0.7237	0.331860
4	1000	0.0275%	0.4996	3.008	-0.6789	-0.7298	0.242645
5	100	2.002%	0.5067	3.030	-0.6763	-0.7293	0.501108
6	200	2.002%	0.5004	3.000	-0.6748	-0.7352	0.419827
7	500	2.002%	0.5029	3.018	-0.6787	-0.7329	0.298438
8	1000	2.002%	0.5009	3.016	-0.6749	-0.7381	0.211292
9	100	4.994%	0.5097	3.094	-0.6844	-0.7352	0.465790
10	200	4.994%	0.5034	3.005	-0.6654	-0.7493	0.359169
11	500	4.994%	0.5027	2.995	-0.6663	-0.7495	0.222824
12	1000	4.994%	0.5009	3.000	-0.6655	-0.7530	0.137920
13	100	9.984%	0.5049	3.058	-0.6548	-0.7808	0.372519
14	200	9.984%	0.5046	3.042	-0.6599	-0.7759	0.257915
15	500	9.984%	0.5015	3.029	-0.6586	-0.7802	0.125261
16	1000	9.984%	0.5022	3.017	-0.6585	-0.7781	0.059100
17	100	25%	0.5086	3.092	-0.6392	-0.8868	0.144363
18	200	25%	0.5030	3.056	-0.6276	-0.9029	0.049851
19	500	25%	0.4999	3.031	-0.6241	-0.9082	0.005885
20	1000	25%	0.5008	2.984	-0.6223	-0.8991	0.001614

[1] $\widehat{\pi}$ is the proportion of the right censored observations. $\widehat{\lambda}^{AFT}$ and $\widehat{k}^{AFT}$ are the estimated scale and shape parameters from the continuous time AFT. $\widehat{\alpha}_0^{WPH}$ and $\widehat{\alpha}_{log(j)}^{WPH}$ are the coefficients of the grouped time PH model. $\widehat{\gamma}_0^{naive}$ is the MLCV naive bandwidth of grouped time.
[2] The number of the Monte Carlo samples is 500 in all simulations.
[3] aMSE for the models in this table are shown in Table D.15 below.

Table D.15: aMSE for the estimation techniques for the unconditional hazard rate with shape=3 and scale=0.50.

	n	$\widehat{\pi}$	aMSE($\widehat{\theta}^{WPH}(j)$)	aMSE($\widehat{\theta}^{SPH}(j)$)	aMSE($\widehat{\theta}^{EKH}(j)$)	aMSE($\widehat{\theta}^{DNK}(j)$)
1	100	0.0275%	0.000340	0.001172	0.001296	0.002039
2	200	0.0275%	0.000269	0.000591	0.000895	0.001484
3	500	0.0275%	0.000228	0.000244	0.000478	0.000819
4	1000	0.0275%	0.000219	0.000128	0.000265	0.000455
5	100	2.002%	0.000359	0.001308	0.001243	0.001860
6	200	2.002%	0.000279	0.000618	0.000844	0.001293
7	500	2.002%	0.000237	0.000264	0.000449	0.000674
8	1000	2.002%	0.000228	0.000136	0.000245	0.000348
9	100	4.994%	0.000398	0.001405	0.001237	0.001665
10	200	4.994%	0.000304	0.000660	0.000756	0.001002
11	500	4.994%	0.000256	0.000293	0.000364	0.000417
12	1000	4.994%	0.000246	0.000154	0.000197	0.000195
13	100	9.984%	0.000469	0.001688	0.001393	0.001335
14	200	9.984%	0.000366	0.000818	0.000833	0.000712
15	500	9.984%	0.000313	0.000344	0.000437	0.000463
16	1000	9.984%	0.000291	0.000199	0.000272	0.001203
17	100	25%	0.000959	0.005564	0.003607	0.018809
18	200	25%	0.000780	0.003545	0.002220	0.029333
19	500	25%	0.000666	0.001579	0.001222	0.038294
20	1000	25%	0.000581	0.001011	0.000852	0.043179

[1] The average mean square error (aMES) is computed for the models:$\widehat{\theta}^{WPH}(j)$, the grouped time hazard with $log(j)$ specification of the elapsed time in the spell variable. $\widehat{\theta}^{SPH}(j)$, the piecewise grouped time hazard. $\widehat{\theta}^{EKH}(j)$ kernel external hazard model, and $\widehat{\theta}^{DNK}(j)$ the density implied hazard.

Table D.16: Coefficients of parametric models for the unconditional hazard for Weibull with shape=3 and scale=0.25.

	n	$\widehat{\pi}$	$\widehat{k}^{AFT}$	$\widehat{\lambda}^{AFT}$	$\widehat{\alpha}_0^{WPH}$	$\widehat{\alpha}_{log(j)}^{WPH}$	$\widehat{\gamma}_0^{naive}$
1	100	0.027%	0.2532	3.303	-0.6171	-0.9983	0.626261
2	200	0.027%	0.2528	3.125	-0.6088	-1.0010	0.499590
3	500	0.027%	0.2507	3.005	-0.6039	-1.0030	0.315253
4	1000	0.027%	0.2504	3.045	-0.6098	-1.0020	0.194329
5	100	1.991%	0.2532	3.178	-0.5712	-1.0360	0.635536
6	200	1.991%	0.2529	3.131	-0.5818	-1.0320	0.499794
7	500	1.991%	0.2504	3.062	-0.5804	-1.0340	0.315184
8	1000	1.991%	0.2502	3.031	-0.5776	-1.0360	0.196037
9	100	4.993%	0.2522	3.202	-0.5417	-1.0800	0.642241
10	200	4.993%	0.2514	3.131	-0.5414	-1.0820	0.505253
11	500	4.993%	0.2508	3.038	-0.5434	-1.0800	0.320692
12	1000	4.993%	0.2500	3.019	-0.5422	-1.0820	0.198012
13	100	9.996%	0.2545	3.224	-0.4915	-1.1510	0.677300
14	200	9.996%	0.2515	3.207	-0.5061	-1.1480	0.527941
15	500	9.996%	0.2505	3.066	-0.4960	-1.1520	0.329701
16	1000	9.996%	0.2505	3.013	-0.4951	-1.1500	0.205204
17	100	25%	0.2543	3.359	-0.4209	-1.3290	0.365517
18	200	25%	0.2517	3.158	-0.4144	-1.3280	0.228685
19	500	25%	0.2500	3.104	-0.4178	-1.3280	0.088246
20	1000	25%	0.2501	3.011	-0.4087	-1.3310	0.023976

[1] $\widehat{\pi}$ is the proportion of the right censored observations. $\widehat{\lambda}^{AFT}$ and $\widehat{k}^{AFT}$ are the estimated scale and shape parameters from the continuous time AFT. $\widehat{\alpha}_0^{WPH}$ and $\widehat{\alpha}_{log(j)}^{WPH}$ are the coefficients of the grouped time PH model. $\widehat{\gamma}_0^{naive}$ is the MLCV naive bandwidth of grouped time.

[2] The number of the Monte Carlo samples is 500 in all simulations.

[3] aMSE for the models in this table are shown in Table D.17 below.

Table D.17: aMSE for the estimation techniques for the unconditional hazard rate with shape=3 and scale=0.25.

	n	$\widehat{\pi}$	aMSE($\widehat{\theta}^{WPH}(j)$)	aMSE($\widehat{\theta}^{SPH}(j)$)	aMSE($\widehat{\theta}^{EKH}(j)$)	aMSE($\widehat{\theta}^{DNK}(j)$)
1	100	0.027%	0.000865	0.000501	0.002035	0.002027
2	200	0.027%	0.000815	0.000243	0.001276	0.001227
3	500	0.027%	0.000792	0.000101	0.000525	0.000493
4	1000	0.027%	0.000782	0.000051	0.000222	0.000214
5	100	1.991%	0.000793	0.000521	0.002075	0.002068
6	200	1.991%	0.000755	0.000253	0.001278	0.001222
7	500	1.991%	0.000722	0.000100	0.000533	0.000503
8	1000	1.991%	0.000713	0.000053	0.000221	0.000214
9	100	4.993%	0.000711	0.000519	0.002110	0.002107
10	200	4.993%	0.000663	0.000261	0.001303	0.001246
11	500	4.993%	0.000639	0.000104	0.000550	0.000514
12	1000	4.993%	0.000628	0.000053	0.000227	0.000219
13	100	9.996%	0.000611	0.000523	0.002342	0.002353
14	200	9.996%	0.000560	0.000266	0.001432	0.001364
15	500	9.996%	0.000534	0.000105	0.000576	0.000536
16	1000	9.996%	0.000527	0.000054	0.000243	0.000237
17	100	25%	0.000443	0.000702	0.000994	0.001199
18	200	25%	0.000399	0.000335	0.000479	0.000935
19	500	25%	0.000370	0.000127	0.000166	0.000805
20	1000	25%	0.000363	0.000065	0.000081	0.000741

[1] The average mean square error (aMES) is computed for the models:$\widehat{\theta}^{WPH}(j)$, the grouped time hazard with $log(j)$ specification of the elapsed time in the spell variable. $\widehat{\theta}^{SPH}(j)$, the piecewise grouped time hazard. $\widehat{\theta}^{EKH}(j)$ kernel external hazard model, and $\widehat{\theta}^{DNK}(j)$ the density implied hazard.

Table D.18: Coefficients of the parametric conditional hazard models with a baseline hazard that follows Weibull with scale=1 and shape=1.

	n	$\widehat{\pi}$	$\widehat{\lambda}^{AFT}$	$\widehat{k}^{AFT}$	$\widehat{\beta}^{*}_{x^c}$	$\widehat{\beta}^{*}_{x^d}$	$\widehat{\beta}^{AFT}_{x^c}$	$\widehat{\beta}^{AFT}_{x^d}$	$\widehat{\alpha}^{AFT}_{log(j)}$	$\widehat{\beta}^{WPH}_{x^c}$	$\widehat{\beta}^{WPH}_{x^d}$	$\widehat{\beta}^{SPH}_{x^c}$	$\widehat{\beta}^{SPH}_{x^d}$	$\widehat{\gamma}_0$	$\widehat{h}_{x^c}$	$\widehat{\gamma}_{x^d}$
1	100	0%	1.0181	1.0341	-0.1418	-0.5007	0.1462	0.5186	-0.0324	0.1469	0.5357	0.1456	0.5312	0.031730	1.663475	0.168386
2	200	0%	1.0068	1.0112	-0.1539	-0.5168	0.1556	0.5224	-0.0103	0.1567	0.5263	0.1560	0.5250	0.016383	1.627709	0.116421
3	500	0%	0.9976	1.0051	-0.1508	-0.4981	0.1516	0.5002	0.0001	0.1532	0.5052	0.1530	0.5041	0.006003	1.229184	0.076301
4	1000	0%	1.0025	1.0023	-0.1502	-0.5004	0.1505	0.5016	-0.0048	0.1511	0.5020	0.1508	0.5012	0.002966	1.023685	0.054520
5	100	2.016%	1.0027	1.0276	-0.1554	-0.4948	0.1600	0.5088	-0.0304	0.1637	0.5135	0.1625	0.5106	0.034550	1.512605	0.184472
6	200	2.095%	1.0018	1.0135	-0.1490	-0.5040	0.1512	0.5107	-0.0283	0.1528	0.5145	0.1521	0.5116	0.015683	1.514930	0.126240
7	500	2.104%	1.0037	1.0090	-0.1517	-0.5015	0.1532	0.5059	-0.0212	0.1504	0.5015	0.1501	0.5011	0.005970	1.230653	0.078527
8	1000	2.088%	0.9997	1.0004	-0.1502	-0.5014	0.1503	0.5016	-0.0185	0.1494	0.4980	0.1492	0.4975	0.002902	1.003755	0.059419
9	100	5.178%	0.9993	1.0287	-0.1508	-0.4908	0.1551	0.5050	-0.0585	0.1534	0.5056	0.1518	0.5029	0.029083	1.596301	0.178568
10	200	5.142%	1.0095	1.0130	-0.1531	-0.5066	0.1546	0.5135	-0.0614	0.1513	0.5039	0.1504	0.5015	0.014885	1.555789	0.115428
11	500	5.222%	1.0044	1.0037	-0.1522	-0.5084	0.1528	0.5105	-0.0489	0.1495	0.4969	0.1494	0.4966	0.005788	1.193136	0.082931
12	1000	5.233%	0.9991	1.0002	-0.1479	-0.5005	0.1479	0.5006	-0.0470	0.1434	0.4869	0.1435	0.4869	0.002857	1.069635	0.063123
13	100	10.04%	1.0107	1.0274	-0.1537	-0.5056	0.1574	0.5194	-0.1040	0.1540	0.5037	0.1535	0.5024	0.027838	1.502260	0.207824
14	200	9.89%	1.0014	1.0126	-0.1504	-0.5063	0.1523	0.5130	-0.0929	0.1429	0.4943	0.1428	0.4940	0.013906	1.463149	0.129072
15	500	9.986%	1.0046	1.0082	-0.1512	-0.5001	0.1526	0.5040	-0.0944	0.1445	0.4837	0.1445	0.4840	0.005493	1.260498	0.088414
16	1000	9.923%	1.0015	1.0022	-0.1515	-0.5018	0.1518	0.5028	-0.0907	0.1449	0.4809	0.1450	0.4812	0.002676	1.017698	0.058934
17	100	25.58%	1.0069	1.0181	-0.1469	-0.5108	0.1491	0.5182	-0.2386	0.1410	0.4640	0.1415	0.4669	5.1×10^{-10}	1.811211	0.228102
18	200	25.45%	0.9975	1.0080	-0.1485	-0.4943	0.1491	0.4971	-0.2253	0.1329	0.4464	0.1335	0.4483	3.8×10^{-10}	1.438381	0.186680
19	500	25.58%	1.0010	1.0043	-0.1488	-0.5029	0.1495	0.5045	-0.2352	0.1314	0.4556	0.1322	0.4587	2.4×10^{-9}	1.111623	0.098031
20	1000	25.61%	0.9991	1.0026	-0.1499	-0.4980	0.1502	0.4992	-0.2353	0.1347	0.4493	0.1355	0.4522	3.6×10^{-9}	0.951499	0.062131

[1] The baseline hazard is Weibull with $\lambda = 1$ and $k = 1$. The conditional transition time is simulated from the function $Y_i^c \sim Weibull(scale = \lambda exp\left[-\frac{1}{k}\left(0.15X_i^c - 0.5X_i^d\right)\right], shape = k)$, the interval length $b = 1$. The censoring time is simulated from $C_i^c \sim Uniform(0, u_\pi)$, where π is the proportion of the right censored observations.

[2] The distribution of the covariates are $X_i^c \sim Normal(0,1)$ and $Pr(X_i^d = 1) = 0.5$.

[3] Average Mean Square Error of the models in this table are shown in Table D.24 below.

Table D.19: Coefficients of the parametric conditional hazard models with a baseline hazard that follows Weibull with scale=1 and shape=0.75.

	n	$\widehat{\pi}$	$\widehat{\lambda}^{AFT}$	$\widehat{k}^{AFT}$	$\widehat{\beta}^*_{x^c}$	$\widehat{\beta}^*_{x^d}$	$\widehat{\beta}^{AFT}_{x^c}$	$\widehat{\beta}^{AFT}_{x^d}$	$\widehat{\alpha}^{AFT}_{log(j)}$	$\widehat{\beta}^{WPH}_{x^c}$	$\widehat{\beta}^{WPH}_{x^d}$	$\widehat{\beta}^{SPH}_{x^c}$	$\widehat{\beta}^{SPH}_{x^d}$	$\widehat{\gamma}_0$	$\widehat{h}_{x^c}$	$\widehat{\gamma}_{x^d}$
1	100	0%	1.0127	0.7704	-0.2119	-0.6680	0.1627	0.5144	-0.0213	0.1694	0.5243	0.1712	0.5256	0.084450	1.746265	0.204636
2	200	0%	1.0136	0.7569	-0.1967	-0.6837	0.1487	0.5177	-0.0212	0.1517	0.5247	0.1524	0.5280	0.051266	1.680770	0.145650
3	500	0%	0.9980	0.7520	-0.1990	-0.6619	0.1497	0.4979	-0.0108	0.1512	0.4994	0.1520	0.5009	0.022133	1.417583	0.105088
4	1000	0%	1.0045	0.7518	-0.2006	-0.6781	0.1508	0.5099	-0.0150	0.1499	0.5108	0.1506	0.5118	0.011650	1.239671	0.073361
5	100	1.96%	1.0113	0.7745	-0.2010	-0.6720	0.1558	0.5202	-0.0323	0.1586	0.5244	0.1597	0.5274	0.077328	1.653344	0.204021
6	200	2.155%	1.0092	0.7617	-0.2080	-0.6715	0.1586	0.5117	-0.0342	0.1569	0.5180	0.1579	0.5204	0.039953	1.665936	0.152365
7	500	2.217%	0.9983	0.7512	-0.2048	-0.6601	0.1538	0.4959	-0.0255	0.1524	0.4966	0.1533	0.4979	0.020375	1.375597	0.109519
8	1000	2.158%	1.0007	0.7508	-0.1998	-0.6715	0.1500	0.5041	-0.0273	0.1488	0.5028	0.1495	0.5039	0.009933	1.165822	0.076162
9	100	5.352%	1.0124	0.7685	-0.1904	-0.6785	0.1465	0.5214	-0.0606	0.1496	0.5314	0.1511	0.5325	0.062205	1.781556	0.193109
10	200	5.37%	1.0031	0.7601	-0.2025	-0.6680	0.1540	0.5079	-0.0438	0.1536	0.5035	0.1545	0.5056	0.037915	1.818313	0.160514
11	500	5.355%	0.9973	0.7529	-0.2011	-0.6639	0.1514	0.4998	-0.0387	0.1498	0.4917	0.1505	0.4931	0.015784	1.459373	0.109617
12	1000	5.371%	1.0005	0.7527	-0.1976	-0.6650	0.1487	0.5006	-0.0425	0.1463	0.4912	0.1467	0.4924	0.008207	1.225351	0.081507
13	100	9.826%	1.0155	0.7686	-0.1932	-0.6806	0.1478	0.5227	-0.0811	0.1491	0.5206	0.1498	0.5240	0.052836	1.772693	0.211110
14	200	9.872%	0.9993	0.7579	-0.2051	-0.6578	0.1555	0.4983	-0.0689	0.1506	0.4851	0.1509	0.4869	0.026380	1.691936	0.182961
15	500	9.904%	1.0012	0.7559	-0.1993	-0.6661	0.1506	0.5032	-0.0703	0.1454	0.4896	0.1459	0.4906	0.005895	1.388940	0.119607
16	1000	9.987%	1.0024	0.7526	-0.2011	-0.6697	0.1514	0.5039	-0.0733	0.1486	0.4901	0.1489	0.4909	0.002746	1.174853	0.082914
17	100	25.79%	1.0351	0.7645	-0.2021	-0.6775	0.1542	0.5147	-0.2056	0.1399	0.4868	0.1407	0.4872	9.5×10^{-10}	1.646474	0.264103
18	200	25.66%	1.0110	0.7597	-0.2042	-0.6564	0.1548	0.4975	-0.1901	0.1441	0.4624	0.1444	0.4622	2.4×10^{-9}	1.358224	0.193854
19	500	25.62%	0.9980	0.7541	-0.2083	-0.6561	0.1571	0.4941	-0.1869	0.1464	0.4615	0.1465	0.4615	3.9×10^{-9}	1.136616	0.103612
20	1000	25.72%	1.0085	0.7519	-0.1984	-0.6754	0.1492	0.5078	-0.1966	0.1380	0.4727	0.1380	0.4728	3.8×10^{-9}	1.046054	0.066073

[1] The baseline hazard is Weibull with $\lambda = 1$ and $k = 0.75$. The conditional transition time is simulated from the function $Y_i^c \sim Weibull(scale = \lambda exp\left[-\frac{1}{k}\left(0.15X_i^c - 0.5X_i^d\right)\right], shape = k)$, the interval length $b = 1$. The censoring time is simulated from $C_i^c \sim Uniform(0, u_\pi)$, where π is the proportion of the right censored observations.

[2] The distribution of the covariates are $X_i^c \sim Normal(0,1)$ and $Pr(X_i^d = 1) = 0.5$.

[3] Average Mean Square Error of the models in this table are shown in Table D.25 below.

Table D.20: Coefficients of the parametric conditional hazard models with a baseline hazard that follows Weibull with scale=1 and shape=0.50.

	n	$\widehat{\pi}$	$\widehat{\lambda}^{AFT}$	$\widehat{k}^{AFT}$	$\widehat{\beta}^{*}_{x^c}$	$\widehat{\beta}^{*}_{x^d}$	$\widehat{\beta}^{AFT}_{x^c}$	$\widehat{\beta}^{AFT}_{x^d}$	$\widehat{\alpha}^{AFT}_{log(j)}$	$\widehat{\beta}^{WPH}_{x^c}$	$\widehat{\beta}^{WPH}_{x^d}$	$\widehat{\beta}^{SPH}_{x^c}$	$\widehat{\beta}^{SPH}_{x^d}$	$\widehat{\gamma}_0$	$\widehat{h}_{x^c}$	$\widehat{\gamma}_{x^d}$
1	100	0%	1.0048	0.5108	-0.2956	-0.9811	0.1513	0.5019	-0.0528	0.1583	0.5216	0.1601	0.5197	0.204512	1.844412	0.209003
2	200	0%	1.0244	0.5062	-0.2967	-1.0122	0.1500	0.5118	-0.0626	0.1515	0.5216	0.1529	0.5227	0.150272	2.228866	0.185913
3	500	0%	1.0000	0.5034	-0.2933	-0.9884	0.1477	0.4977	-0.0474	0.1477	0.5017	0.1486	0.5025	0.090054	1.861803	0.150580
4	1000	0%	1.0068	0.5013	-0.3023	-1.0025	0.1516	0.5025	-0.0474	0.1513	0.5020	0.1520	0.5027	0.058052	1.527618	0.115567
5	100	2.34%	1.0389	0.5113	-0.3004	-1.0109	0.1547	0.5162	-0.0674	0.1589	0.5375	0.1595	0.5366	0.179932	2.027212	0.243760
6	200	2.359%	1.0065	0.5046	-0.2940	-1.0043	0.1482	0.5063	-0.0383	0.1504	0.5068	0.1519	0.5078	0.122173	2.054896	0.217545
7	500	2.318%	0.9972	0.5036	-0.3042	-0.9912	0.1531	0.4992	-0.0477	0.1526	0.5007	0.1532	0.5012	0.067951	1.790861	0.155898
8	1000	2.335%	1.0028	0.5009	-0.3025	-0.9998	0.1515	0.5009	-0.0482	0.1521	0.5019	0.1526	0.5021	0.040083	1.492282	0.120326
9	100	5.892%	1.0527	0.5128	-0.2847	-1.0209	0.1456	0.5219	-0.0688	0.1514	0.5357	0.1517	0.5352	0.136166	2.041293	0.242211
10	200	5.643%	1.0201	0.5061	-0.2946	-1.0007	0.1488	0.5058	-0.0537	0.1485	0.5078	0.1485	0.5106	0.085495	2.249848	0.218610
11	500	5.714%	0.9935	0.5013	-0.2983	-0.9908	0.1494	0.4966	-0.0441	0.1485	0.4957	0.1488	0.4961	0.035250	1.974115	0.168862
12	1000	5.641%	0.9993	0.5019	-0.3027	-0.9995	0.1519	0.5016	-0.0490	0.1519	0.5003	0.1520	0.5006	0.014727	1.431191	0.125216
13	100	10.02%	1.0370	0.5104	-0.3000	-1.0100	0.1533	0.5158	-0.0637	0.1589	0.5176	0.1599	0.5199	0.089546	1.915178	0.270866
14	200	10.01%	1.0283	0.5051	-0.3049	-1.0252	0.1538	0.5177	-0.0731	0.1558	0.5218	0.1555	0.5234	0.041404	2.066512	0.213151
15	500	9.903%	1.0065	0.5044	-0.3006	-1.0003	0.1515	0.5043	-0.0578	0.1520	0.5012	0.1522	0.5018	0.011472	1.637559	0.163606
16	1000	9.962%	0.9979	0.5006	-0.3035	-0.9943	0.1518	0.4976	-0.0545	0.1507	0.4948	0.1507	0.4950	0.002848	1.427207	0.117944
17	100	25.36%	1.0606	0.5140	-0.2808	-1.0317	0.1431	0.5252	-0.1452	0.1370	0.5148	0.1368	0.5155	6.6×10^{-9}	1.708673	0.270214
18	200	25.71%	1.0318	0.5036	-0.3147	-1.0080	0.1583	0.5059	-0.1343	0.1507	0.4824	0.1507	0.4827	1.1×10^{-9}	1.456940	0.194341
19	500	25.63%	1.0127	0.5027	-0.3014	-1.0000	0.1511	0.5021	-0.1338	0.1435	0.4801	0.1435	0.4803	3.5×10^{-10}	1.404432	0.126084
20	1000	25.59%	1.0096	0.5000	-0.3037	-1.0107	0.1517	0.5050	-0.1348	0.1463	0.4837	0.1464	0.4837	7.6×10^{-10}	1.050624	0.082009

[1] The baseline hazard is Weibull with $\lambda = 1$ and $k = 0.50$. The conditional transition time is simulated from the function $Y_i^c \sim Weibull(scale = \lambda exp\left[-\frac{1}{k}\left(0.15X_i^c - 0.5X_i^d\right)\right], shape = k)$, the interval length $b = 1$. The censoring time is simulated from $C_i^c \sim Uniform(0, u_\pi)$, where π is the proportion of the right censored observations.

[2] The distribution of the covariates are $X_i^c \sim Normal(0,1)$ and $Pr(X_i^d = 1) = 0.5$.

[3] Average Mean Square Error of the models in this table are shown in Table D.26 below.

Table D.21: Coefficients of the parametric conditional hazard models with a baseline hazard that follows Weibull with scale=3 and shape=0.25.

	n	$\widehat{\pi}$	$\widehat{\lambda}^{AFT}$	$\widehat{k}^{AFT}$	$\widehat{\beta}^{*}_{x^c}$	$\widehat{\beta}^{*}_{x^d}$	$\widehat{\beta}^{AFT}_{x^c}$	$\widehat{\beta}^{AFT}_{x^d}$	$\widehat{\alpha}^{AFT}_{log(j)}$	$\widehat{\beta}^{WPH}_{x^c}$	$\widehat{\beta}^{WPH}_{x^d}$	$\widehat{\beta}^{SPH}_{x^c}$	$\widehat{\beta}^{SPH}_{x^d}$	$\widehat{\gamma}_0$	$\widehat{h}_{x^c}$	$\widehat{\gamma}_{x^d}$
1	100	0%	1.1029	0.2548	-0.6054	-1.9954	0.1543	0.5091	-0.1813	0.1647	0.5413	0.1600	0.5199	0.643972	2.411807	0.280893
2	200	0%	1.0950	0.2535	-0.6007	-2.0188	0.1522	0.5118	-0.1899	0.1587	0.5336	0.1541	0.5180	0.470423	2.751627	0.261261
3	500	0%	1.0476	0.2521	-0.6033	-1.9874	0.1523	0.5012	-0.1906	0.1556	0.5207	0.1514	0.5037	0.276687	2.365626	0.259052
4	1000	0%	1.0138	0.2508	-0.6008	-2.0029	0.1508	0.5024	-0.1811	0.1540	0.5184	0.1495	0.5006	0.169619	2.267491	0.234998
5	100	2.538%	1.1445	0.2555	-0.5962	-2.0352	0.1523	0.5193	-0.1499	0.1632	0.5452	0.1595	0.5348	0.641270	2.680651	0.247452
6	200	2.596%	1.0535	0.2521	-0.6140	-1.9901	0.1544	0.5016	-0.1396	0.1617	0.5243	0.1582	0.5118	0.501364	2.613850	0.260571
7	500	2.53%	0.9972	0.2513	-0.5892	-1.9793	0.1480	0.4972	-0.1376	0.1524	0.5124	0.1488	0.5010	0.293055	2.741476	0.251872
8	1000	2.626%	1.0174	0.2507	-0.5967	-2.0037	0.1496	0.5023	-0.1463	0.1539	0.5170	0.1506	0.5050	0.173233	2.321664	0.229360
9	100	5.594%	1.1608	0.2544	-0.5884	-1.9609	0.1493	0.4974	-0.1133	0.1611	0.5278	0.1581	0.5170	0.463039	2.520369	0.290195
10	200	5.6%	1.1014	0.2536	-0.6220	-2.0313	0.1572	0.5150	-0.1229	0.1644	0.5299	0.1615	0.5217	0.320118	2.840678	0.280789
11	500	5.652%	1.0324	0.2512	-0.5893	-2.0038	0.1480	0.5029	-0.1192	0.1513	0.5188	0.1479	0.5084	0.168991	2.677933	0.272659
12	1000	5.64%	1.0236	0.2505	-0.6034	-2.0069	0.1511	0.5027	-0.1199	0.1545	0.5158	0.1513	0.5049	0.091631	2.424865	0.240351
13	100	10.1%	1.2495	0.2569	-0.6200	-1.9810	0.1587	0.5082	-0.1001	0.1670	0.5300	0.1650	0.5237	0.252606	2.584472	0.314539
14	200	10.01%	1.0973	0.2541	-0.6040	-1.9707	0.1534	0.4999	-0.0836	0.1574	0.5123	0.1549	0.5040	0.144475	3.035465	0.321272
15	500	10.02%	1.0376	0.2514	-0.5891	-2.0087	0.1480	0.5046	-0.0823	0.1513	0.5124	0.1483	0.5045	0.055919	2.644324	0.260832
16	1000	9.986%	1.0152	0.2503	-0.6013	-2.0040	0.1505	0.5015	-0.0839	0.1539	0.5109	0.1511	0.5026	0.018958	2.236561	0.226331
17	100	19.42%	1.2585	0.2547	-0.6426	-2.0595	0.1632	0.5238	-0.0558	0.1669	0.5300	0.1664	0.5270	0.047647	2.409321	0.333052
18	200	19.53%	1.1372	0.2505	-0.6158	-2.0347	0.1538	0.5087	-0.0599	0.1553	0.5126	0.1547	0.5092	1.3×10^{-8}	2.667402	0.280982
19	500	19.42%	1.0520	0.2511	-0.6105	-2.0094	0.1532	0.5040	-0.0535	0.1532	0.5061	0.1523	0.5036	1.2×10^{-10}	1.745418	0.177750
20	1000	19.3%	1.0067	0.2509	-0.5933	-1.9823	0.1487	0.4971	-0.0493	0.1503	0.4997	0.1494	0.4967	1.5×10^{-10}	1.370596	0.116227

[1] The baseline hazard is Weibull with $\lambda = 1$ and $k = 0.25$. The conditional transition time is simulated from the function $Y_i^c \sim Weibull(scale = \lambda exp\left[-\frac{1}{k}\left(0.15X_i^c - 0.5X_i^d\right)\right], shape = k)$, the interval length $b = 1$. The censoring time is simulated from $C_i^c \sim Uniform(0, u_\pi)$, where π is the proportion of the right censored observations.

[2] The distribution of the covariates are $X_i^c \sim Normal(0,1)$ and $Pr(X_i^d = 1) = 0.5$.

[3] Average Mean Square Error of the models in this table are shown in Table D.27 below.

Table D.22: Coefficients of the parametric conditional hazard models with a baseline hazard that follows Weibull with scale=3 and shape=1.

	n	$\widehat{\pi}$	$\widehat{\lambda}^{AFT}$	$\widehat{k}^{AFT}$	$\widehat{\beta}^{*}_{x^c}$	$\widehat{\beta}^{*}_{x^d}$	$\widehat{\beta}^{AFT}_{x^c}$	$\widehat{\beta}^{AFT}_{x^d}$	$\widehat{\alpha}^{AFT}_{log(j)}$	$\widehat{\beta}^{WPH}_{x^c}$	$\widehat{\beta}^{WPH}_{x^d}$	$\widehat{\beta}^{SPH}_{x^c}$	$\widehat{\beta}^{SPH}_{x^d}$	$\widehat{\gamma}_0$	$\widehat{h}_{x^c}$	$\widehat{\gamma}_{x^d}$
1	100	0%	2.9761	1.0232	-0.1493	-0.4925	0.1533	0.5045	-1.1165	0.1538	0.5069	0.1532	0.5041	0.214758	2.309527	0.262196
2	200	0%	3.0043	1.0144	-0.1524	-0.4999	0.1547	0.5076	-1.1195	0.1563	0.5110	0.1562	0.5104	0.131337	1.991616	0.200574
3	500	0%	2.9959	1.0070	-0.1503	-0.5019	0.1512	0.5053	-1.1023	0.1508	0.5051	0.1506	0.5048	0.062278	1.589427	0.134119
4	1000	0%	3.0121	1.0029	-0.1488	-0.5043	0.1492	0.5058	-1.1074	0.1497	0.5065	0.1496	0.5066	0.031399	1.348343	0.098535
5	100	1.972%	2.9996	1.0296	-0.1636	-0.4979	0.1688	0.5125	-1.1435	0.1716	0.5191	0.1704	0.5147	0.195638	2.134901	0.269904
6	200	2.01%	2.9992	1.0142	-0.1529	-0.5043	0.1551	0.5115	-1.1142	0.1558	0.5111	0.1555	0.5090	0.125092	2.062059	0.217018
7	500	2.053%	2.9994	1.0067	-0.1505	-0.4969	0.1515	0.5000	-1.1120	0.1520	0.5009	0.1520	0.5008	0.057517	1.606949	0.144038
8	1000	2.067%	3.0016	1.0016	-0.1487	-0.5003	0.1488	0.5011	-1.1068	0.1491	0.5014	0.1490	0.5010	0.029923	1.410518	0.101966
9	100	5.1%	3.0150	1.0256	-0.1456	-0.5071	0.1496	0.5184	-1.1505	0.1510	0.5223	0.1498	0.5189	0.199216	2.408939	0.269524
10	200	5.236%	2.9910	1.0096	-0.1468	-0.5009	0.1481	0.5059	-1.1191	0.1485	0.5082	0.1485	0.5055	0.128277	2.043909	0.214221
11	500	5.287%	3.0049	1.0033	-0.1500	-0.5060	0.1504	0.5074	-1.1189	0.1506	0.5075	0.1508	0.5078	0.053006	1.659980	0.143529
12	1000	5.199%	3.0086	1.0021	-0.1488	-0.5024	0.1491	0.5034	-1.1161	0.1490	0.5036	0.1488	0.5033	0.028356	1.436319	0.104091
13	100	9.998%	3.0113	1.0309	-0.1535	-0.5100	0.1579	0.5252	-1.1612	0.1604	0.5279	0.1593	0.5242	0.168890	2.511227	0.269927
14	200	9.893%	3.0075	1.0095	-0.1544	-0.5006	0.1560	0.5055	-1.1350	0.1551	0.5050	0.1550	0.5041	0.100921	2.115604	0.216676
15	500	9.894%	3.0096	1.0047	-0.1504	-0.5025	0.1511	0.5050	-1.1299	0.1503	0.5041	0.1501	0.5041	0.045444	1.696586	0.147583
16	1000	9.916%	2.9895	1.0030	-0.1493	-0.4960	0.1497	0.4975	-1.1230	0.1495	0.4966	0.1494	0.4963	0.020225	1.429432	0.108159
17	100	25.68%	3.0386	1.0282	-0.1439	-0.5039	0.1471	0.5160	-1.1911	0.1459	0.5139	0.1454	0.5105	0.070163	2.399050	0.335554
18	200	25.5%	2.9939	1.0120	-0.1536	-0.4992	0.1550	0.5045	-1.1635	0.1534	0.5028	0.1534	0.5031	0.027444	2.186393	0.268487
19	500	25.7%	3.0192	1.0067	-0.1473	-0.5007	0.1481	0.5035	-1.1640	0.1472	0.4980	0.1473	0.4981	9.5×10^{-9}	1.830930	0.182319
20	1000	25.62%	3.0005	1.0041	-0.1476	-0.4985	0.1481	0.5005	-1.1596	0.1475	0.4974	0.1477	0.4979	1.8×10^{-8}	1.366262	0.115949

[1] The baseline hazard is Weibull with $\lambda = 3$ and $k = 1$. The conditional transition time is simulated from the function $Y_i^c \sim Weibull(scale = \lambda exp\left[-\frac{1}{k}\left(0.15X_i^c - 0.5X_i^d\right)\right], shape = k)$, the interval length $b = 1$. The censoring time is simulated from $C_i^c \sim Uniform(0, u_\pi)$, where π is the proportion of the right censored observations.

[2] The distribution of the covariates are $X_i^c \sim Normal(0,1)$ and $Pr(X_i^d = 1) = 0.5$.

[3] Average Mean Square Error of the models in this table are shown in Table D.28 below.

Table D.23: Coefficients of parametric conditional hazard models with a baseline hazard that follows Weibull with scale=3 and shape=0.25.

	n	$\widehat{\pi}$	$\widehat{\lambda}^{AFT}$	$\widehat{k}^{AFT}$	$\widehat{\beta}^{*}_{x^c}$	$\widehat{\beta}^{*}_{x^d}$	$\widehat{\beta}^{AFT}_{x^c}$	$\widehat{\beta}^{AFT}_{x^d}$	$\widehat{\alpha}^{AFT}_{log(j)}$	$\widehat{\beta}^{WPH}_{x^c}$	$\widehat{\beta}^{WPH}_{x^d}$	$\widehat{\beta}^{SPH}_{x^c}$	$\widehat{\beta}^{SPH}_{x^d}$	$\widehat{\gamma}_0$	$\widehat{h}_{x^c}$	$\widehat{\gamma}_{x^d}$
1	100	0%	3.4042	0.2568	-0.6328	-1.9601	0.1624	0.5039	-0.5475	0.1671	0.5256	0.1610	0.5104	0.624521	2.755378	0.326451
2	200	0%	3.1383	0.2532	-0.6062	-2.0037	0.1538	0.5069	-0.5262	0.1578	0.5115	0.1512	0.4997	0.478477	2.664145	0.300243
3	500	0%	3.1266	0.2514	-0.6170	-1.9889	0.1551	0.4999	-0.5350	0.1554	0.5026	0.1503	0.4908	0.297994	2.634167	0.294745
4	1000	0%	3.0058	0.2509	-0.6033	-1.9753	0.1513	0.4956	-0.5239	0.1512	0.4971	0.1468	0.4850	0.185564	2.519149	0.279747
5	100	2.624%	3.5428	0.2573	-0.6360	-2.0115	0.1636	0.5169	-0.5140	0.1703	0.5333	0.1662	0.5260	0.642140	2.577694	0.314285
6	200	2.687%	3.2194	0.2532	-0.6171	-2.0409	0.1563	0.5170	-0.4904	0.1610	0.5265	0.1577	0.5176	0.481480	3.037504	0.316282
7	500	2.641%	3.0363	0.2507	-0.5982	-1.9922	0.1499	0.4992	-0.4911	0.1536	0.5093	0.1502	0.4992	0.300618	2.749155	0.303677
8	1000	2.605%	3.0558	0.2508	-0.6082	-1.9995	0.1525	0.5013	-0.4991	0.1564	0.5117	0.1528	0.5016	0.186951	2.495890	0.279900
9	100	5.786%	3.6762	0.2559	-0.6162	-2.0318	0.1576	0.5183	-0.4800	0.1660	0.5336	0.1624	0.5291	0.675991	2.783750	0.319404
10	200	5.503%	3.2326	0.2533	-0.6157	-1.9994	0.1558	0.5059	-0.4667	0.1603	0.5202	0.1569	0.5130	0.514072	2.949421	0.312479
11	500	5.582%	3.1018	0.2515	-0.6079	-2.0121	0.1530	0.5057	-0.4656	0.1590	0.5185	0.1559	0.5113	0.309487	2.876076	0.295469
12	1000	5.622%	3.0459	0.2502	-0.5860	-2.0062	0.1467	0.5018	-0.4594	0.1502	0.5125	0.1477	0.5047	0.189746	2.647244	0.276807
13	100	10.21%	3.9181	0.2547	-0.6641	-2.1026	0.1688	0.5325	-0.4471	0.1748	0.5508	0.1713	0.5453	0.502498	3.089833	0.341944
14	200	10.05%	3.3368	0.2535	-0.6276	-2.0405	0.1588	0.5167	-0.4307	0.1650	0.5288	0.1616	0.5208	0.355461	2.830189	0.329708
15	500	10.02%	3.1470	0.2511	-0.6083	-2.0012	0.1526	0.5021	-0.4273	0.1569	0.5132	0.1537	0.5040	0.189176	3.086601	0.301280
16	1000	9.875%	3.0481	0.2511	-0.5928	-1.9919	0.1489	0.5000	-0.4256	0.1513	0.5110	0.1482	0.5016	0.100751	2.647107	0.280887
17	100	19.56%	3.8732	0.2539	-0.6298	-2.0230	0.1594	0.5114	-0.3664	0.1637	0.5118	0.1619	0.5094	0.188427	3.162805	0.368211
18	200	19.2%	3.2578	0.2532	-0.5991	-2.0166	0.1513	0.5095	-0.3531	0.1559	0.5170	0.1540	0.5116	0.089062	2.862950	0.337845
19	500	19.28%	3.1622	0.2512	-0.5931	-2.0305	0.1490	0.5095	-0.3627	0.1512	0.5148	0.1493	0.5086	0.014542	2.670095	0.292607
20	1000	19.44%	3.0825	0.2502	-0.6073	-2.0134	0.1519	0.5036	-0.3603	0.1535	0.5093	0.1515	0.5028	3.2×10^{-9}	1.863720	0.215105

[1] The baseline hazard is Weibull with $\lambda = 3$ and $k = 0.25$. The conditional transition time is simulated from the function $Y_i^c \sim Weibull(scale = \lambda exp\left[-\frac{1}{k}\left(0.15X_i^c - 0.5X_i^d\right)\right], shape = k)$, the interval length $b = 1$. The censoring time is simulated from $C_i^c \sim Uniform(0, u_\pi)$, where π is the proportion of the right censored observations.

[2] The distribution of the covariates are $X_i^c \sim Normal(0, 1)$ and $Pr(X_i^d = 1) = 0.5$.

[3] Average Mean Square Error of the models in this table are shown in Table D.31 below.

Table D.24: aMSE of the conditional hazard models with a baseline hazard that follows Weibull with scale=1 and shape=1.

		aMSE of the baseline hazard				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.009422	0.005871	0.018917	0.017679	0.015362
200	0%	0.004979	0.002956	0.012469	0.012521	0.011051
500	0%	0.002128	0.001263	0.004942	0.004852	0.004765
1000	0%	0.000552	0.000602	0.002546	0.002535	0.002520
100	2%	0.009388	0.005858	0.021351	0.018620	0.015286
200	2.1%	0.004785	0.002892	0.012459	0.011116	0.011174
500	2.1%	0.001986	0.001254	0.005912	0.005767	0.005619
1000	2.1%	0.001038	0.000663	0.003224	0.003179	0.002881
100	5.2%	0.010517	0.006876	0.023322	0.019302	0.015143
200	5.1%	0.005435	0.003332	0.015000	0.013359	0.013017
500	5.2%	0.002559	0.001688	0.007607	0.007329	0.006348
1000	5.2%	0.001502	0.000986	0.003741	0.003666	0.003551
100	10%	0.013367	0.008985	0.026657	0.023195	0.016888
200	9.9%	0.007647	0.004681	0.019568	0.018403	0.014387
500	8.6%	0.004131	0.002814	0.010714	0.010365	0.010736
1000	2.3%	0.002787	0.002090	0.007172	0.007176	0.007721
100	25.6%	0.048616	0.031028	0.071970	0.067462	0.025686
200	25.4%	0.040132	0.026190	0.059442	0.055053	0.026912
500	25.6%	0.031770	0.021063	0.061396	0.058058	0.022023
1000	25.6%	0.028069	0.019197	0.065133	0.061038	0.021735

		aMSE of the conditional hazard at 8 combination of values for x^c and x^d				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.009084	0.006252	0.024328	0.018806	0.019269
200	0%	0.004936	0.003248	0.017757	0.015530	0.015283
500	0%	0.002078	0.001301	0.008489	0.008061	0.008505
1000	0%	0.001028	0.000656	0.004912	0.004866	0.005027
100	2%	0.009146	0.006252	0.027768	0.021603	0.020048
200	2.1%	0.004841	0.003209	0.018014	0.014885	0.014214
500	2.1%	0.001315	0.001322	0.010138	0.009507	0.008885
1000	2.1%	0.001095	0.000741	0.005916	0.005702	0.004938
100	5.2%	0.010413	0.007188	0.031669	0.022320	0.018518
200	5.1%	0.005724	0.003784	0.022529	0.018387	0.015113
500	5.2%	0.002717	0.001879	0.012897	0.011923	0.008851
1000	5.2%	0.001597	0.001098	0.007754	0.007577	0.005412
100	10%	0.013484	0.009357	0.042079	0.032420	0.020626
200	9.9%	0.008126	0.005210	0.031282	0.024810	0.014650
500	8.6%	0.004366	0.003054	0.018315	0.017031	0.010750
1000	2.3%	0.002826	0.002210	0.013339	0.013123	0.007205
100	25.6%	0.054042	0.033099	0.105069	0.095835	0.027076
200	25.4%	0.045419	0.028980	0.091311	0.082136	0.025214
500	25.6%	0.035238	0.022891	0.087015	0.080321	0.021288
1000	25.6%	0.031325	0.021175	0.087927	0.079459	0.019038

[1] The point of (X^c, X^d) that used to calculated the average aMSE are $(0,0), (1,0), (-1,0), (\overline{X}^c, 0), (0,1), (1,1), (-1,1), (\overline{X}^c, 1)$. The first point measures the hazard of the baseline group.

[2] The interval length is $b = 1$ and the number of intervals that used in the aMSE is up to the interval that include the 99^{th} quantile of Y_i^c.

Table D.25: aMSE of the conditional hazard models with a baseline hazard that follows Weibull with scale=1 and shape=0.75.

		aMSE of the baseline hazard				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.006940	0.048827	0.058289	0.046248	0.044857
200	0%	0.003287	0.027438	0.034146	0.029698	0.028670
500	0%	0.001555	0.011063	0.013884	0.013553	0.013078
1000	0%	0.000977	0.004987	0.006436	0.006325	0.006296
100	2%	0.007295	0.052714	0.064164	0.052219	0.047440
200	2.2%	0.003610	0.032012	0.036962	0.032454	0.030902
500	2.2%	0.001866	0.012594	0.016169	0.015279	0.015602
1000	2.2%	0.001165	0.006431	0.008338	0.008046	0.008244
100	5.4%	0.008974	0.062216	0.070903	0.056331	0.053284
200	5.4%	0.005175	0.037865	0.044981	0.039783	0.038655
500	5.4%	0.002663	0.016169	0.020787	0.019659	0.021187
1000	5.4%	0.001712	0.008047	0.010448	0.010243	0.011985
100	9.8%	0.011786	0.069181	0.078031	0.065597	0.062131
200	9.9%	0.007334	0.049783	0.057436	0.052652	0.049521
500	0.4%	0.004575	0.027439	0.033357	0.031091	0.041666
1000	8.7%	0.003680	0.016286	0.021365	0.020705	0.036943
100	25.8%	0.030013	0.076369	0.084180	0.075951	0.170736
200	25.7%	0.023410	0.064482	0.067797	0.061643	0.168171
500	25.6%	0.016921	0.057889	0.058982	0.055198	0.164336
1000	25.7%	0.019395	0.055069	0.056669	0.054179	0.163617

		aMSE of the conditional hazard at 8 combination of values for x^c and x^d				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.008294	0.056832	0.073041	0.055712	0.054970
200	0%	0.005068	0.032695	0.047845	0.040034	0.041812
500	0%	0.002355	0.013068	0.023319	0.022035	0.024242
1000	0%	0.001401	0.006177	0.014209	0.013722	0.015813
100	2%	0.010353	0.060595	0.078766	0.062114	0.055979
200	2.2%	0.005383	0.037006	0.050043	0.043367	0.042618
500	2.2%	0.002755	0.014872	0.026172	0.024293	0.025727
1000	2.2%	0.001635	0.007892	0.017776	0.017269	0.017529
100	5.4%	0.012414	0.071346	0.085857	0.067072	0.058779
200	5.4%	0.007404	0.044257	0.059396	0.051679	0.046700
500	5.4%	0.003956	0.019754	0.032826	0.030717	0.029049
1000	5.4%	0.002592	0.009740	0.020774	0.020144	0.019053
100	9.8%	0.016602	0.085458	0.093154	0.082931	0.062021
200	9.9%	0.010443	0.061012	0.076296	0.069410	0.050221
500	0.4%	0.006393	0.034418	0.049328	0.046499	0.033773
1000	8.7%	0.005094	0.020675	0.034719	0.033597	0.032686
100	25.8%	0.036503	0.105658	0.124436	0.112756	0.126385
200	25.7%	0.031214	0.092443	0.106040	0.096821	0.123262
500	25.6%	0.026957	0.083456	0.092790	0.086586	0.120842
1000	25.7%	0.025420	0.079057	0.086036	0.081751	0.120313

[1] The point of (X^c, X^d) that used to calculated the average aMSE are $(0,0), (1,0), (-1,0), (\overline{X}^c, 0), (0,1), (1,1), (-1,1), (\overline{X}^c, 1)$. The first point measures the hazard of the baseline group.

[2] The interval length is $b = 1$ and the number of intervals that used in the aMSE is up to the interval that include the 99^{th} quantile of Y_i^c.

Table D.26: aMSE of the conditional hazard models with a baseline hazard that follows Weibull with scale=1 and shape=0.50.

		aMSE of the baseline hazard				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.002420	0.026135	0.026722	0.027219	0.020988
200	0%	0.001411	0.015114	0.015411	0.015553	0.009285
500	0%	0.001140	0.005605	0.006018	0.006222	0.003737
1000	0%	0.001106	0.001702	0.002036	0.002114	0.001806
100	2.3%	0.002697	0.026321	0.027389	0.024944	0.019330
200	2.4%	0.001975	0.017042	0.017724	0.017611	0.011668
500	2.3%	0.001357	0.007145	0.007857	0.007892	0.006087
1000	2.3%	0.001331	0.002506	0.003055	0.003099	0.002799
100	5.9%	0.003388	0.031097	0.031233	0.026431	0.026202
200	5.6%	0.002479	0.021135	0.022524	0.020370	0.018557
500	5.7%	0.001690	0.010561	0.011452	0.011099	0.013877
1000	5.6%	0.001828	0.008247	0.008758	0.008483	0.010831
100	10%	0.004318	0.034285	0.037190	0.032731	0.055136
200	10%	0.003012	0.020722	0.022958	0.021570	0.044942
500	0.3%	0.002490	0.011545	0.012626	0.012076	0.042039
1000	6.2%	0.002402	0.005836	0.007264	0.007234	0.042615
100	25.4%	0.013267	0.031949	0.039338	0.034544	0.267160
200	25.7%	0.010811	0.023583	0.025930	0.023037	0.260776
500	25.6%	0.009338	0.017586	0.018737	0.018325	0.254356
1000	25.6%	0.009146	0.015158	0.015878	0.015878	0.253064

		aMSE of the conditional hazard at 8 combination of values for x^c and x^d				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.004712	0.035411	0.031844	0.031383	0.036859
200	0%	0.002553	0.020972	0.019401	0.020563	0.021165
500	0%	0.001862	0.008200	0.009272	0.009529	0.009638
1000	0%	0.001772	0.002804	0.004864	0.004982	0.005197
100	2.3%	0.005012	0.036403	0.032555	0.029482	0.036351
200	2.4%	0.003258	0.023011	0.021892	0.021537	0.020967
500	2.3%	0.002273	0.010408	0.010896	0.010930	0.009460
1000	2.3%	0.002178	0.003090	0.006029	0.006007	0.005490
100	5.9%	0.005936	0.042144	0.037471	0.032537	0.042432
200	5.6%	0.004176	0.027882	0.027736	0.025425	0.022608
500	5.7%	0.003384	0.014953	0.015617	0.015299	0.015359
1000	5.6%	0.003056	0.011124	0.012308	0.011956	0.011497
100	10%	0.007403	0.045039	0.045586	0.040274	0.057366
200	10%	0.004000	0.028405	0.030091	0.028286	0.043980
500	0.3%	0.004094	0.016699	0.018569	0.018121	0.037339
1000	6.2%	0.003962	0.008905	0.012943	0.012880	0.037277
100	25.4%	0.019470	0.046321	0.058244	0.052513	0.223783
200	25.7%	0.016522	0.036056	0.044515	0.041744	0.218340
500	25.6%	0.014355	0.028028	0.033978	0.033535	0.213037
1000	25.6%	0.013946	0.024239	0.029164	0.029164	0.211720

[1] The point of (X^c, X^d) that used to calculated the average aMSE are $(0,0), (1,0), (-1,0), (\overline{X}^c, 0), (0,1), (1,1), (-1,1), (\overline{X}^c, 1)$. The first point measures the hazard of the baseline group.

[2] The interval length is $b = 1$ and the number of intervals that used in the aMSE is up to the interval that include the 99^{th} quantile of Y_i^c.

Table D.27: aMSE of the conditional hazard models with a baseline hazard that follows Weibull with scale=1 and shape=0.25.

		aMSE of the baseline hazard				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.001085	0.001464	0.001961	0.007025	0.003209
200	0%	0.000954	0.000669	0.000873	0.003449	0.001588
500	0%	0.000883	0.000250	0.000353	0.001394	0.000677
1000	0%	0.000860	0.000123	0.000192	0.000608	0.000336
100	2.5%	0.000944	0.001661	0.002112	0.006526	0.003317
200	2.6%	0.000799	0.000675	0.000896	0.003864	0.001767
500	2.5%	0.000734	0.000263	0.000384	0.001603	0.000774
1000	2.6%	0.000698	0.000128	0.000203	0.000647	0.000376
100	5.6%	0.000815	0.001838	0.002326	0.005061	0.003299
200	5.6%	0.000681	0.000801	0.001020	0.002396	0.001502
500	5.7%	0.000605	0.000278	0.000397	0.000802	0.000689
1000	5.6%	0.000587	0.000135	0.000212	0.000322	0.000393
100	10.1%	0.000844	0.003396	0.003930	0.004768	0.014769
200	10%	0.000689	0.001478	0.001829	0.002087	0.011880
500	10%	0.000606	0.000654	0.000855	0.000854	0.013422
1000	8.6%	0.000581	0.000352	0.000453	0.000441	0.014233
100	19.4%	0.002273	0.009135	0.012439	0.012186	0.202182
200	19.5%	0.001664	0.005743	0.007064	0.006814	0.189427
500	19.4%	0.001340	0.001943	0.002629	0.002626	0.176944
1000	19.3%	0.001208	0.001012	0.001383	0.001383	0.172556

		aMSE of the conditional hazard at 8 combination of values for x^c and x^d				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.001948	0.002684	0.003639	0.011986	0.008136
200	0%	0.001575	0.001213	0.001522	0.006011	0.003129
500	0%	0.001387	0.000444	0.000615	0.002237	0.001035
1000	0%	0.001337	0.000219	0.000390	0.001072	0.000580
100	2.5%	0.001657	0.002754	0.004054	0.012246	0.006290
200	2.6%	0.001294	0.001212	0.001611	0.006377	0.003362
500	2.5%	0.001137	0.000465	0.000661	0.002572	0.001130
1000	2.6%	0.001083	0.000228	0.000409	0.001139	0.000595
100	5.6%	0.001380	0.003340	0.003611	0.008407	0.009282
200	5.6%	0.001123	0.001473	0.001675	0.003891	0.003635
500	5.7%	0.000960	0.000504	0.000662	0.001291	0.000866
1000	5.6%	0.000922	0.000242	0.000416	0.000607	0.000486
100	10.1%	0.001453	0.005791	0.005089	0.006517	0.018723
200	10%	0.001110	0.002473	0.002394	0.002853	0.011942
500	10%	0.000972	0.001057	0.001230	0.001244	0.012958
1000	8.6%	0.000934	0.000556	0.000745	0.000730	0.013777
100	19.4%	0.003784	0.014095	0.015046	0.014631	0.194868
200	19.5%	0.002708	0.008453	0.009356	0.008154	0.182390
500	19.4%	0.002161	0.003317	0.004547	0.004550	0.171084
1000	19.3%	0.002016	0.001766	0.002962	0.002962	0.167159

[1] The point of (X^c, X^d) that used to calculated the average aMSE are $(0,0), (1,0), (-1,0), (\overline{X}^c, 0), (0,1), (1,1), (-1,1), (\overline{X}^c, 1)$. The first point measures the hazard of the baseline group.

[2] The interval length is $b = 1$ and the number of intervals that used in the aMSE is up to the interval that include the 99^{th} quantile of Y_i^c.

Table D.28: aMSE of the conditional hazard models with a baseline hazard that follows Weibull with scale=3 and shape=1.

		aMSE of the baseline hazard				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.001773	0.002090	0.004233	0.004014	0.003289
200	0%	0.000858	0.001019	0.002108	0.002007	0.001832
500	0%	0.000336	0.000385	0.000818	0.000761	0.000733
1000	0%	0.000156	0.000185	0.000372	0.000348	0.000345
100	2%	0.001804	0.001867	0.003625	0.003471	0.003201
200	2%	0.000838	0.000968	0.002008	0.001884	0.002014
500	2.1%	0.000336	0.000397	0.000762	0.000716	0.000878
1000	2.1%	0.000168	0.000197	0.000382	0.000357	0.000503
100	5.1%	0.001979	0.002034	0.004024	0.003842	0.004269
200	5.2%	0.000904	0.001064	0.002124	0.002054	0.002751
500	5.3%	0.000334	0.000398	0.000745	0.000695	0.001331
1000	5.1%	0.000178	0.000210	0.000376	0.000353	0.000877
100	9.8%	0.002179	0.002016	0.004157	0.004110	0.006191
200	9.9%	0.000925	0.000972	0.001946	0.001870	0.003980
500	9.9%	0.000397	0.000455	0.000731	0.000685	0.002446
1000	1.6%	0.000194	0.000208	0.000372	0.000358	0.002027
100	25.7%	0.002677	0.002405	0.004871	0.004712	0.021505
200	25.5%	0.001470	0.001318	0.002119	0.002075	0.018886
500	25.7%	0.000925	0.000766	0.000856	0.000849	0.014735
1000	25.6%	0.000653	0.000536	0.000513	0.000511	0.013241

		aMSE of the conditional hazard at 8 combination of values for x^c and x^d				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.003339	0.003432	0.006576	0.007159	0.006117
200	0%	0.001471	0.001577	0.003614	0.003779	0.003614
500	0%	0.000561	0.000592	0.001869	0.001843	0.001879
1000	0%	0.000258	0.000292	0.001197	0.001175	0.001203
100	2%	0.003578	0.003368	0.006677	0.007380	0.006145
200	2%	0.001474	0.001570	0.003745	0.003812	0.003565
500	2.1%	0.000562	0.000618	0.002022	0.001493	0.001907
1000	2.1%	0.000276	0.000305	0.001294	0.001273	0.001208
100	5.1%	0.003619	0.003496	0.007197	0.007954	0.006274
200	5.2%	0.001565	0.001666	0.004192	0.004216	0.003830
500	5.3%	0.000596	0.000654	0.002132	0.002087	0.001909
1000	5.1%	0.000296	0.000327	0.001363	0.001341	0.001260
100	9.8%	0.003803	0.003490	0.006986	0.007451	0.006662
200	9.9%	0.001614	0.001637	0.004269	0.004282	0.004067
500	9.9%	0.000635	0.000676	0.002336	0.002284	0.002288
1000	1.6%	0.000347	0.000350	0.001526	0.001507	0.001735
100	25.7%	0.004538	0.004064	0.009851	0.010146	0.015438
200	25.5%	0.002508	0.002190	0.006277	0.006176	0.012768
500	25.7%	0.001450	0.001214	0.003864	0.003848	0.009806
1000	25.6%	0.001017	0.000830	0.002579	0.002574	0.009195

[1] The point of (X^c, X^d) that used to calculated the average aMSE are $(0,0), (1,0), (-1,0), (\overline{X}^c, 0), (0,1), (1,1), (-1,1), (\overline{X}^c, 1)$. The first point measures the hazard of the baseline group.

[2] The interval length is $b = 1$ and the number of intervals that used in the aMSE is up to the interval that include the 99^{th} quantile of Y_i^c.

Table D.29: aMSE of the conditional hazard models with a baseline hazard that follows Weibull with scale=3 and shape=0.75.

		aMSE of the baseline hazard				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.001694	0.005676	0.008143	0.006798	0.005532
200	0%	0.000803	0.002355	0.003637	0.003131	0.002778
500	0%	0.000319	0.000967	0.001554	0.001361	0.001280
1000	0%	0.000206	0.000477	0.000807	0.000743	0.000714
100	2.2%	0.001541	0.005497	0.008132	0.007031	0.005891
200	2.1%	0.000737	0.002682	0.003989	0.003395	0.003334
500	2.2%	0.000311	0.000947	0.001519	0.001365	0.001456
1000	2.1%	0.000193	0.000468	0.000797	0.000728	0.000817
100	5.5%	0.001711	0.005732	0.008745	0.007172	0.007256
200	5.5%	0.000718	0.002484	0.003811	0.003250	0.003772
500	5.5%	0.000333	0.000990	0.001563	0.001408	0.002060
1000	5.4%	0.000210	0.000502	0.000823	0.000756	0.001329
100	10%	0.001793	0.006454	0.009105	0.007630	0.009758
200	9.9%	0.000801	0.002887	0.004133	0.003642	0.005811
500	9.8%	0.000362	0.001158	0.001765	0.001617	0.003623
1000	9.4%	0.000232	0.000553	0.000847	0.000803	0.002485
100	25.9%	0.002728	0.011895	0.016292	0.014843	0.059273
200	25.6%	0.001444	0.004881	0.006624	0.006343	0.047432
500	25.7%	0.001094	0.002149	0.002906	0.002885	0.040385
1000	25.7%	0.000903	0.001331	0.001744	0.001741	0.037771

		aMSE of the conditional hazard at 8 combination of values for x^c and x^d				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.003341	0.009332	0.013343	0.012625	0.010603
200	0%	0.001373	0.003686	0.006712	0.006148	0.005648
500	0%	0.000555	0.001550	0.003304	0.003113	0.003030
1000	0%	0.000350	0.000764	0.001979	0.001916	0.001895
100	2.2%	0.003203	0.009145	0.013748	0.012460	0.011264
200	2.1%	0.001445	0.004348	0.006917	0.006440	0.005787
500	2.2%	0.000555	0.001516	0.003299	0.003129	0.002915
1000	2.1%	0.000336	0.000761	0.002096	0.002020	0.001866
100	5.5%	0.003175	0.009109	0.013774	0.012160	0.011140
200	5.5%	0.001303	0.004186	0.007372	0.006537	0.006503
500	5.5%	0.000573	0.001586	0.003488	0.003312	0.003083
1000	5.4%	0.000349	0.000791	0.002191	0.002109	0.001971
100	10%	0.003217	0.010190	0.014259	0.012421	0.013020
200	9.9%	0.001459	0.004631	0.007407	0.006826	0.007012
500	9.8%	0.000621	0.001783	0.003923	0.003750	0.003860
1000	9.4%	0.000390	0.000893	0.002436	0.002374	0.002499
100	25.9%	0.004646	0.018034	0.022379	0.020905	0.050521
200	25.6%	0.002571	0.007859	0.012088	0.011810	0.038542
500	25.7%	0.001821	0.003502	0.006851	0.006828	0.032688
1000	25.7%	0.001433	0.002094	0.004582	0.004578	0.030969

[1] The point of (X^c, X^d) that used to calculated the average aMSE are $(0,0), (1,0), (-1,0), (\overline{X}^c, 0), (0,1), (1,1), (-1,1), (\overline{X}^c, 1)$. The first point measures the hazard of the baseline group.

[2] The interval length is $b = 1$ and the number of intervals that used in the aMSE is up to the interval that include the 99^{th} quantile of Y_i^c.

Table D.30: aMSE of the conditional hazard models with a baseline hazard that follows Weibull with scale=3 and shape=0.50.

		aMSE of the baseline hazard				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.000792	0.003291	0.004309	0.005524	0.003567
200	0%	0.000546	0.001326	0.001886	0.002709	0.001846
500	0%	0.000398	0.000480	0.000745	0.001254	0.000867
1000	0%	0.000357	0.000239	0.000395	0.000677	0.000475
100	2.5%	0.000866	0.004288	0.005299	0.006047	0.004711
200	2.3%	0.000543	0.001441	0.002024	0.002673	0.001972
500	2.4%	0.000401	0.000525	0.000788	0.001049	0.000880
1000	2.4%	0.000355	0.000258	0.000409	0.000515	0.000470
100	5.9%	0.000844	0.004387	0.005309	0.005751	0.004584
200	5.7%	0.000546	0.001678	0.002197	0.002449	0.002423
500	5.7%	0.000413	0.000560	0.000869	0.000880	0.001170
1000	5.7%	0.000366	0.000281	0.000456	0.000433	0.000767
100	10.1%	0.000872	0.006254	0.007167	0.007027	0.008507
200	2.7%	0.000594	0.002602	0.003228	0.003287	0.004801
500	4%	0.000462	0.000708	0.001041	0.000322	0.002484
1000	10%	0.000425	0.000346	0.000546	0.000521	0.002016
100	25.5%	0.002001	0.009486	0.012651	0.011660	0.121134
200	25.5%	0.001335	0.005021	0.006488	0.006387	0.125883
500	25.6%	0.001014	0.001804	0.002365	0.002365	0.128161
1000	25.5%	0.000890	0.000988	0.001379	0.001379	0.125455

		aMSE of the conditional hazard at 8 combination of values for x^c and x^d				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.001549	0.005728	0.006985	0.011295	0.008368
200	0%	0.000917	0.002274	0.003457	0.005491	0.004436
500	0%	0.000627	0.000840	0.001581	0.002546	0.001865
1000	0%	0.000552	0.000417	0.001028	0.001562	0.001200
100	2.5%	0.001775	0.007231	0.007549	0.010635	0.009254
200	2.3%	0.000944	0.002501	0.003557	0.005026	0.004579
500	2.4%	0.000640	0.000914	0.001675	0.002191	0.001720
1000	2.4%	0.000550	0.000450	0.001109	0.001359	0.001077
100	5.9%	0.001620	0.007384	0.007583	0.009221	0.009813
200	5.7%	0.000950	0.002897	0.003607	0.004317	0.004213
500	5.7%	0.000681	0.000984	0.001832	0.001938	0.001638
1000	5.7%	0.000587	0.000492	0.001179	0.001193	0.001033
100	10.1%	0.001688	0.009874	0.009109	0.009583	0.012244
200	2.7%	0.001029	0.004385	0.004488	0.004652	0.005046
500	4%	0.000779	0.001253	0.002006	0.001980	0.002343
1000	10%	0.000702	0.000603	0.001360	0.001341	0.001741
100	25.5%	0.003391	0.014590	0.016039	0.015222	0.111126
200	25.5%	0.002196	0.007740	0.009154	0.008711	0.114156
500	25.6%	0.001709	0.003085	0.004748	0.004748	0.117009
1000	25.5%	0.001494	0.001693	0.003231	0.003231	0.116500

[1] The point of (X^c, X^d) that used to calculated the average aMSE are $(0,0), (1,0), (-1,0), (\overline{X}^c, 0), (0,1), (1,1), (-1,1), (\overline{X}^c, 1)$. The first point measures the hazard of the baseline group.

[2] The interval length is $b = 1$ and the number of intervals that used in the aMSE is up to the interval that include the 99^{th} quantile of Y_i^c.

Table D.31: aMSE of the conditional hazard models with a baseline hazard that follows Weibull with scale=3 and shape=0.25.

		aMSE of the baseline hazard				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.000887	0.000681	0.000954	0.003033	0.001732
200	0%	0.000785	0.000321	0.000498	0.001855	0.000977
500	0%	0.000729	0.000125	0.000224	0.000790	0.000425
1000	0%	0.000712	0.000064	0.000146	0.000344	0.000212
100	2.6%	0.000766	0.000671	0.000952	0.003277	0.001819
200	2.7%	0.000670	0.000324	0.000501	0.001906	0.001012
500	2.6%	0.000617	0.000128	0.000239	0.000817	0.000446
1000	2.6%	0.000599	0.000064	0.000142	0.000348	0.000217
100	5.8%	0.000668	0.000704	0.001006	0.003453	0.002050
200	5.5%	0.000573	0.000321	0.000495	0.002079	0.001146
500	5.6%	0.000523	0.000129	0.000233	0.000849	0.000488
1000	5.6%	0.000505	0.000068	0.000151	0.000363	0.000249
100	10.2%	0.000579	0.000775	0.001056	0.002391	0.001658
200	10.1%	0.000480	0.000362	0.000538	0.001281	0.000943
500	10%	0.000431	0.000133	0.000240	0.000443	0.000474
1000	9.9%	0.000420	0.000068	0.000149	0.000180	0.000322
100	19.6%	0.000790	0.002506	0.003132	0.003491	0.022341
200	19.2%	0.000626	0.001082	0.001549	0.001474	0.022047
500	19.3%	0.000534	0.000393	0.000627	0.000607	0.025909
1000	19.4%	0.000496	0.000191	0.000333	0.000332	0.029336

		aMSE of the conditional hazard at 8 combination of values for x^c and x^d				
n	$\widehat{\pi}$	$\widehat{\theta}^{CWPH}$	$\widehat{\theta}^{CSPH}$	$\widehat{\theta}^{BKH}$	$\widehat{\theta}^{CEKH}$	$\widehat{\theta}^{CDKH}$
100	0%	0.001519	0.001243	0.001577	0.005025	0.004329
200	0%	0.001264	0.000567	0.000718	0.002893	0.001664
500	0%	0.001113	0.000213	0.000358	0.001323	0.000757
1000	0%	0.001079	0.000110	0.000250	0.000656	0.000416
100	2.6%	0.001327	0.001263	0.001560	0.005471	0.003754
200	2.7%	0.001091	0.000578	0.000711	0.002901	0.001705
500	2.6%	0.000960	0.000223	0.000369	0.001358	0.000765
1000	2.6%	0.000927	0.000111	0.000251	0.000671	0.000420
100	5.8%	0.001120	0.001206	0.001572	0.005695	0.003870
200	5.5%	0.000931	0.000570	0.000725	0.003226	0.001904
500	5.6%	0.000832	0.000231	0.000373	0.001420	0.000790
1000	5.6%	0.000789	0.000118	0.000265	0.000688	0.000425
100	10.2%	0.000972	0.001419	0.001564	0.004128	0.004101
200	10.1%	0.000788	0.000656	0.000749	0.001986	0.001291
500	10%	0.000682	0.000235	0.000381	0.000791	0.000609
1000	9.9%	0.000662	0.000119	0.000264	0.000379	0.000370
100	19.6%	0.001272	0.004231	0.003955	0.004301	0.023218
200	19.2%	0.001026	0.001963	0.001670	0.002073	0.021333
500	19.3%	0.000867	0.000700	0.000936	0.000928	0.024820
1000	19.4%	0.000812	0.000342	0.000592	0.000592	0.028527

[1] The point of (X^c, X^d) that used to calculated the average aMSE are $(0,0), (1,0), (-1,0), (\overline{X}^c, 0), (0,1), (1,1), (-1,1), (\overline{X}^c, 1)$. The first point measures the hazard of the baseline group.

[2] The interval length is $b = 1$ and the number of intervals that used in the aMSE is up to the interval that include the 99^{th} quantile of Y_i^c.