REGRESSION ON MEDIAN RESIDUAL LIFE FUNCTION FOR CENSORED SURVIVAL DATA

by

Hanna Bandos

M.S., V.N. Karazin Kharkiv National University, 2000

Submitted to the Graduate Faculty of

The Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Hanna Bandos

It was defended on

July 26, 2007

and approved by

Dissertation Advisor:
Jong-Hyeon Jeong, PhD
Associate Professor
Biostatistics
Graduate School of Public Health
University of Pittsburgh

Dissertation Co-Advisor:
Joseph P. Costantino, DrPH
Professor
Biostatistics
Graduate School of Public Health
University of Pittsburgh

Janice S. Dorman, MS, PhD
Professor
Health Promotion & Development
School of Nursing
University of Pittsburgh

Howard E. Rockette, PhD
Professor
Biostatistics
Graduate School of Public Health
University of Pittsburgh

REGRESSION ON MEDIAN RESIDUAL LIFE FUNCTION FOR CENSORED SURVIVAL DATA

Hanna Bandos, PhD

University of Pittsburgh, 2007

In the analysis of time-to-event data, the *median* residual life (MERL) function has been promoted by many researchers as a practically relevant summary of the residual life distribution. Formally the MERL function at a time point is defined as the median of the remaining lifetimes among survivors beyond that particular time point. Despite its widely recognized usefulness, there is no commonly accepted approach to model the median residual life function.

In this dissertation we introduce two novel regression techniques that model the relationship between the MERL function and covariates of interest at multiple time points simultaneously; proportional median residual life model and accelerated median residual life model. These models have a conceptual similarity to the well-known proportional hazards and accelerated failure time (AFT) models. Inference procedures that we propose for these models permit the data to be right censored.

For the semiparametric analysis under the proportional MERL model, we propose an estimating equation for the regression coefficients. The bootstrap resampling technique is utilized to evaluate the standard errors of the regression coefficient estimates. A simulation study is performed to investigate the proposed inferential approach. The developed method is applied to a real data example from a breast cancer study conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP).

We also propose parametric and semiparametric (under the AFT assumption) inference procedures under the accelerated MERL model. The maximum likelihood inference is

considered for the parametric inference and the Buckley and James method is used to estimate the median residual lifetimes semiparametrically under the AFT assumption. A simulation study is performed to validate the proposed maximum likelihood inference procedure. A generated dataset is used to illustrate statistical analysis via both estimation approaches.

It is very important from a public health perspective to be able to identify the risk factors for a specific disease or condition. The regression techniques presented in this work enable researchers to identify the patients' characteristics that affect their survival experience and describe advantages of a preventive or therapeutic intervention by means of median residual life function in a clinically relevant and intuitively appealing way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my dissertation advisor, Dr. Jong-Hyeon Jeong, for his supervision and assistance through the whole process of preparing of this dissertation. His advises always guided me in the right directions and I consider myself fortunate to have him as an advisor.

I would also like to thank my committee, Dr. Joseph Costantino, Dr. Janice Dorman and Dr. Howard Rockette, for their time and support. My special thanks are to Dr. Costantino – for being my academic advisor and mentor, for his support and encouragement through my graduate studies. He has always been there to listen and give advice.

I am grateful to the Department of Biostatistics for giving me an opportunity to be a graduate student at this outstanding department, to all professors for their inspiring courses which greatly contributed to my professional development. I truly believe that these five years on this program were among the most exciting years of my life.

Finally I would like to thank my family and friends for their love and support.

1.0 INTRODUCTION

Because of the nature of the survival analysis, it is important to have ability to describe or predict the residual life distribution of the patients under study. Even though the simple mean is the most commonly used index to summarize a distribution, the quantiles, including the simple median, are also very useful summary statistics to characterize the survival experience of patients. The mean residual life function (MRL function) and the quantile (median) residual life function (MERL function) are the functional counterparts of these indices commonly used for the time-to-event data. Even though the mean residual life function uniquely defines the lifetime distribution and has many good properties, it still has a number of limitations. When censored observations are present in the sample, the mean residual life function is difficult to be estimated reliably. Moreover, even in case of complete data, the estimated MRL function can be very unstable due to its heavy dependence on the outliers. Due to these facts a better behaving median residual life function has been recommended by many authors to be used for the inferential purposes. Also, compared to a simple median statistic, MERL function as a function of time provides a continuous summary of the residual life distribution. Formally it is defined as $\theta(t) = median(T - t \mid T \geq t)$, where T is a continuous random variable, and it determines the median of the remaining lifetimes among survivors beyond time *t*.

From the practical standpoint, the median residual life function allows researchers and clinicians understand advantages of a particular therapy in terms of the remaining lifetimes of

patients. On the contrary, other well known statistical characteristics that are commonly used in practice for the analysis of survival data, such as hazard function, require a substantial understanding of the statistical concepts.

Often comparison of two or more groups of patients while adjusting for covariates of interest is of great importance, which requires a regression technique to be used. To our knowledge, not many regression techniques exist in the literature for the median residual life function. Those available methods regress the MERL function on important covariates at some specific time point (Ying, Jung and Wei, 1995; McKeague, Subramanian, and Sun, 2001; Yin and Cai, 2005; Jeong, Jung and Bandos, 2007), are focused on a specific class of parametric distributions (Rao, Damaraju, and Alhumoud, 1993), or model the MERL function induced by the accelerated failure time assumption using the Bayesian approach (Gelfand and Kottas, 2003). We propose to develop two kinds of more general frequentist regression methods that could model the relationship between the MERL function and covariates of interest at multiple time points simultaneously – proportional median residual life model and accelerated median residual life model.

1.1 PROPORTIONAL MEDIAN RESIDUAL LIFE MODEL

The proportional median residual life model $\theta(t \mid \boldsymbol{X}_i) = \theta_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}_i)$ by its analytical form resembles the Cox (Cox, 1972) proportional hazards model. Similarly to the Cox model, where the proportionality of the hazard functions regarded as constant over time, this new model specifies that proportionality of the median residual life functions is also constant over time. For the simplest case of the regression with one binary predictor, for example treatment group vs.

control group, our proposed model would indicate that median residual life functions of the control group and treated group are respectively $\theta_0(t)$ and $\eta\theta_0(t)$, where $\eta = e^{\beta}$ and β is the corresponding regression parameter. The positive value of the parameter estimate would indicate an increase of the MERL function for treated patients and value of $\eta = e^{\beta}$ would imply the magnitude of the increase within an interval of interest. On the other hand, a negative value of the parameter estimate would indicate a decrease in the MERL function for treated patients and hence a negative effect of the therapy.

In the proportional median residual life model section we describe the estimation procedure to obtain the parameter estimates and their standard errors. By carrying out the simulation studies we investigate the probability of type I error and perform power analysis over different scenarios. We also apply the new regression technique to a real dataset from a breast cancer trial that was performed by the National Surgical Adjuvant Breast and Bowel Project (NSABP). We state advantages and limitations of our model in the discussion subsection.

1.2 ACCELERATED MEDIAN RESIDUAL LIFE MODEL

The analytical form $\theta(t \mid \boldsymbol{X}_i) = \exp(\boldsymbol{\beta}'\boldsymbol{X}_i)\theta_0(t\exp(-\boldsymbol{\beta}'\boldsymbol{X}_i))$ of the accelerated median residual life model is similar to the accelerated failure time model. For the simplest case of the regression with one binary predictor, for example treatment group vs. control group, this model would indicate that median residual life functions of the control group and treated group are respectively $\theta_0(t)$ and $\eta\theta_0(t/\eta)$, where $\eta = e^{\beta}$ and β is the corresponding regression parameter.

The positive estimate of the regression coefficient would indicate an increase of the MERL function for treated patients during the time period under the study with a shift in the time axis.

In the accelerated median residual life model section we demonstrate that some families of parametric distribution possess the property of uniqueness of one-to-one correspondence between the median residual life function and survival function under the accelerated MERL model and use this fact to introduce a parametric regression. We perform the numerical studies based on the Weibull distribution and report the results. We also describe how semiparametric methods can be used for this type of model under the assumption of the accelerated failure time. We use one of the simulated datasets to illustrate these two techniques for data analysis. We state advantages and limitations of our model in the discussion subsection.

In the second chapter of this dissertation we introduce the median residual life function, its definition and properties and give the overview of the regression techniques that are used in survival analysis. Third and fourth chapters are dedicated to introduction of the proportional MERL model and the accelerated MERL model respectively. Future research directions are outlined in the conclusion section.

2.0 BACKGROUND

2.1 MEDIAN RESIDUAL LIFE FUNCTION

2.1.1 Overview

The simple mean and median are the most commonly used statistics to summarize the center of a distribution. For time-to-event data the functional analogs of these indices exist – the mean residual life function (MRL function) and the median residual life function (MERL function). The mean residual life function uniquely defines the lifetime distribution. Although it has many good properties, it still has a number of limitations. When censored observations are present in the sample, the mean residual life function is difficult to be estimated reliably. Moreover, even in case of complete data, the estimated MRL function can be very unstable due to its heavy dependence on the outliers. Also there are some cases it may not exist (gamma mixture of exponentials where the shape of the gamma distribution is less than 1 (Johnson and Kotz, 1970)). Due to these facts a better behaving median residual life function has been recommended by many authors to be used for the inferential purposes.

A more general concept of the α-percentile residual life function was originally introduced by Haines and Singpurwalla (1974). One of the major difficulties in making inferences based on the percentile residual function is a non-uniqueness of the corresponding life

distribution. This problem has been intensively explored by many authors (Schmittlein and Morrison, 1981; Arnold and Brockett, 1983; Joe and Proschan, 1984; Joe, 1985; Song and Cho, 1995 and Lillo, 2005). Gupta and Langford (1984) under mild assumptions determined a general form of distribution when its median residual life function is known. Ghosh and Mustafi (1986), Csörgö and Csörgö (1987) and Alam and Kulasekera (1993) are among authors who investigated large sample estimation of the MERL function and stochastic properties of such estimators. Also substantial amount of work was done on the confidence bands for the percentile residual life function (Barabas et al., 1986; Aly, 1992; Chung, 1989; Csörgö and Viharos, 1992). Two-sample comparison of the MERL functions is considered in Jeong, Jung and Costantino (2007). There appears to be only a few attempts to develop or describe a regression model for the residual life function (Rao, Damaraju, and Alhumoud, 1993; Gelfand and Kottas, 2003; Jeong, Jung and Bandos, 2007).

2.1.2 Definition and properties

Let $T \geq 0$ be a continuous random variable with the survival function $S(t)$, then we define the median residual life function as the median of the remaining lifetimes among survivors beyond time t or more formally as $\theta(t) = median(T - t \mid T \geq t)$. In other words it can be defined as the length of the interval from time point *t* to the time where one-half of the individuals alive at time *t* will still be alive (Klein and Moeschberger, 2003). This statistic is easily calculated at time point *t* in the presence of censored observations as long as censoring proportion is less than 50% among those who survived up to time point *t*. It is not very sensitive to the skewed distributions. Lastly the MERL function is always finite and is easily obtainable in the closed form for the

distributions with the survival functions available in the closed form. Using the definition of the simple median

$$P(T - t \geq \theta(t) \mid T > t) = \frac{1}{2}$$

$$\frac{P(T - t \geq \theta(t), T > t)}{P(T > t)} = \frac{1}{2}$$

$$\frac{P(T - t \geq \theta(t))}{P(T > t)} = \frac{1}{2}$$

$$\frac{S(t + \theta(t))}{S(t)} = \frac{1}{2}$$

and therefore

$$S(t + \theta(t)) = \frac{1}{2} S(t).$$

If $S(t)$ is strictly decreasing, then the median residual life function can be uniquely defined as

$$\theta(t) = S^{-1}\left[\frac{1}{2} S(t)\right] - t \tag{2.1}$$

The following are some of the basic properties of the median residual life function:

a) $\theta(t) \geq 0$, and $\theta(0) = median(T)$;

b) $\psi(t) = S^{-1}(\frac{1}{2} S(t)) = \theta(t) + t$ is always nondecreasing. It maps $[0, \infty)$ into itself and satisfies the condition $\psi(t) \geq t$ for every $t > 0$;

c) Median residual life function does not uniquely define the underlying distribution.

2.1.3 Estimation of the MERL function

Assuming a specific form of the distribution which has a closed form of its survival function, the median residual life function can be easily calculated using equation (2.1). Below the MERL functions along with the survival functions are calculated for several well-known distributions that are most commonly used in survival analysis.

a) Exponential distribution $S(t) = e^{-\lambda t}$ $\theta(t) = \ln 2 / \lambda$

b) Weibull distribution $S(t) = e^{-(\lambda t)^k}$ $\theta(t) = \frac{1}{\lambda}(\ln 2 + (\lambda t)^k)^{1/k} - t$

c) Pareto distribution $S(t) = (\lambda / t)^{\kappa}$ $\theta(t) = (2^{1/\kappa} - 1)t$

d) Exponential power distribution $S(t) = \exp[1 - e^{(\lambda t)^k}]$ $\theta(t) = \frac{1}{\lambda}\{\ln(\ln 2 + e^{(\lambda t)^k})\}^{1/k} - t$

These formulas allow for parametric estimation of the MERL function using the maximum likelihood estimation technique.

For the nonparametric estimation of the median residual function non-censored and censored cases should be presented separately. First we introduce the notations which are used throughout our work. Let T_i defines failure time for the i^{th} patient in a sample of size n. Because of early termination of study or loss to follow-up, all T_i's may not be completely observed. We define C_i as censoring time for a patient i. Then, for a patient i we observe a pair of variables $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \le C_i)$, where an indicator function $I(\varpi \in W) = 1$ if $\varpi \in W$ and equals to 0 if $\varpi \notin W$. For the complete sample case $Y_i = T_i$ and $\delta_i = 1$ for all $i = 1,..,n$.

For the complete sample case Csörgö and Csörgö (1987) introduced the empirical estimator of the (*1-p*)-percentile residual life function in terms of the empirical estimator of the cumulative distribution (CDF) and sample quantile functions. The median residual life function

estimator is a special case of this estimator, when p = 1/2. The same estimator of the MERL function can be rewritten by incorporating the empirical CDF and its generalized inverse in the following form (Ghosh and Mustafi, 1986; Feng and Kulasekera, 1991):

$$\hat{R}_n(t) = F_n^{-1}(1 - \tfrac{1}{2} S_n(t)) - t \tag{2.2}$$

$$\text{where} \quad S_n(t) = 1 - F_n(t) \quad \text{and} \quad F_n^{-1}(y) = \inf\{x : F_n(x) \geq y\},\ 0 \leq y < 1.$$

For the censored data case, Chung (1989) proposed the *(1-p)*-percentile residual lifetime estimator, which is an analog of the Csörgö and Csörgö (1987) estimator for the complete data. The author used the same form of the estimator, where the empirical CDF is substituted by the Kaplan-Meier (Kaplan and Meier, 1958) product limit estimator and empirical quantile function is substituted by the product limit estimator of the quantile function. Feng and Kulasekera (1991) also rewrote this estimator in the same manner as the equation (2.2) in terms of the Kaplan-Meier CDF and its generalized inverse. We will be using the latter form of the median residual life function estimator and for brevity put it as follows

$$\hat{\theta}(t) = \hat{S}^{-1}\left[\frac{1}{2}\hat{S}(t)\right] - t \tag{2.3}$$

where $\hat{\theta}(t)$ is a MERL function estimator, $\hat{S}(t)$ is the Kaplan-Meier estimator of the survival function and $\hat{S}^{-1}(y) = \inf\{t : \hat{S}(t) \leq y\}$ is the generalized inverse of the Kaplan-Meier estimator. This formula is a straight implication of the equation (2.1). Feng and Kulasekera (1991) also introduced a smooth nonparametric estimator for the percentile residual life function using a kernel type estimator of the CDF for the complete and censored data.

The Kaplan-Meier estimator of the survival function is well defined for all time points less than the largest observed time on study. If the last observation in the sample is censored, estimation of the survival function is a widely recognized challenge in survival analysis. Several

nonparametric methods exist in the literature to address this issue. We choose to estimate the survival function after the last event by the estimate of the survival function at the time of the last event as it was proposed by Gill (1980). Based on the small sample properties of the resulting estimator Klein (1991) showed that this method of estimation is preferable compared to estimating the survival as zero after the last observation in the sample (Efron, 1967), although it still leads to a positively biased estimator. From the equation (2.3) it is clear that the median residual life function estimator is heavily dependent upon the properties of the Kaplan-Meier estimator and therefore an interval where the MERL function can be reliably estimated depend upon the particular sample.

The choice of the estimator of the survival function determines a specific range where the estimator of the median residual life function can be meaningfully defined. While Efron's definition allows for estimation of the MERL function for the entire follow-up period, Gill's definition, which we adopted here, limits the range of estimation to an open interval

$$[0,T) = \left[0, \sup\left\{t \in [0, Y_{(n)}] : \hat{S}(t) \geq 2\hat{S}(Y_{(n)})\right\}\right). \tag{2.4}$$

If the last observation in the sample $Y_{(n)}$ is an event then $T = Y_{(n)}$ and MERL function can be properly estimated on a closed interval [0, *T*].

2.2 REGRESSION MODELS ON SURVIVAL DATA

The regression technique is a useful statistical tool for comparing two or more groups of subjects adjusting for other covariates of interest. There are several regression models available in survival analysis. The Cox proportional hazards model (Cox, 1972) and the accelerated failure

time (AFT) model originally introduced by Miller (1976) are two most commonly used techniques to model censored survival data. Regression model for the simple median was recently introduced by Ying et al. (1995) and presents a novel alternative approach to the regression analysis of survival data.

2.2.1 Cox proportional hazards model

The Cox proportional hazards model is, perhaps, one of the most commonly used regression techniques for time-to-event data. Its fundamental structure is represented in the following form:

$$h(t) = \rho h_0(t) \quad \text{or} \quad S(t) = S_0(t)^{\rho}$$

where ρ is expressed as $\exp(\boldsymbol{\theta}'\mathbf{X})$, vector $\mathbf{X}$ is a vector of patient's covariates, $h(t)$ ($S(t)$) is the hazard (survival) function associated with $\mathbf{X}$, and $\boldsymbol{\theta}$ is a vector of regression parameters. The Cox proportional hazards model does not make any assumptions about the nature or shape of the baseline hazard (survival) function, i.e. $h_0(t)$ ($S_0(t)$) is unspecified. Inferences on the parameter estimates are based on the partial likelihood (Cox, 1972; Cox, 1975) and regression parameters are estimated as those maximizing the partial likelihood function. Algorithms for estimating the Cox regression parameters are available in almost every statistical package.

2.2.2 Accelerated failure time model

The accelerated failure time model presents an alternative to the Cox model and is analogous to the regular linear regression for the noncensored data. It linearly relates the logarithm of survival time to the explanatory variables. The analytical form of this model is as follows

$$\log(T) = \mu + \boldsymbol{\alpha}'\mathbf{X} + \sigma W \ ,$$

where $\boldsymbol{\alpha}$ is the regression parameters and $\boldsymbol{X}$ is the vector of covariates. The choice for the error distribution W determines the distribution for survival times. The model was originally introduced by Miller (1976). If we define $S_0(t)$ as the survival function of the random variable $T_0 = \exp(\mu + \sigma W)$ (the baseline, defined by the set of covariates $\boldsymbol{X} = \boldsymbol{0}$), then the survival function for the random variable T, $S(t)$, will be related to the $S_0(t)$ through the parameter $\rho = \exp(-\boldsymbol{\alpha}'\mathbf{X}) = \exp(\boldsymbol{\gamma}'\mathbf{X})$ as

$$S(t) = S_0(\rho t) \quad \text{or equivalently} \quad h(t) = \rho h_0(\rho t) \, .$$

The accelerated failure time model is often used in the parametric setting, when the error term is assumed to follow a distribution that determines the survival distribution. When the real life data has a baseline that is difficult to fit with a parametric distribution, semiparametric methods for parameter estimation are preferred.

2.2.3 Regression model for a simple median and other related techniques

Regression model for the simple median originally introduced by Ying, Jung and Wei (1995) may be considered as a semiparametric analog of the accelerated failure time model, as it linearly relates the median of failure times (the mean of failure times for the AFT model) to covariates. According to the authors, the main reasons for introducing such model were difficulties associated with estimation of the intercept parameter in the AFT model, simplicity of the median as a measure of centrality, and relatively strong assumptions of the identical distribution of the error terms for estimation and inference procedures for the AFT model. By the first property of the median residual life function defined in section 2.1.2, the simple median can be regarded as

the MERL function at time point 0. Also if $\mathbf{X}_i$ denotes a vector of explanatory variables, $\boldsymbol{\beta}$ denotes a vector of regression parameters (including an intercept) and $\theta(0 \mid \mathbf{X}_i)$ denotes the median of the conditional distribution of $T \mid \mathbf{X}_i$, the regression expression for the simple median model is as follows

$$\theta(0 \mid \mathbf{X}_i) = \boldsymbol{\beta}'\mathbf{X}_i . \tag{2.5}$$

A special type of estimating equation, which is a modification of the least absolute deviations (LAD) method, is used for obtaining the regression estimator.

As before, let T_i and C_i denote failure and censoring time for the i^{th} patient respectively in a sample of size n. Then, for a patient i we observe a pair of variables $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \le C_i)$. For the noncensored case, the LAD estimator for $\boldsymbol{\beta}$ in the model (2.5) is obtained by minimizing $\sum_{i=1}^{n} | T_i - \boldsymbol{\beta}'\mathbf{X}_i |$, which is equivalent to solving the equation

$$\mathbf{U}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i \left\{ I(T_i - \boldsymbol{\beta}'\mathbf{X}_i \ge 0) - \frac{1}{2} \right\} = 0 . \tag{2.6}$$

For the censored case, where Y_i is observed instead of T_i, equation (2.6) is substituted for

$$\mathbf{S}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i \left[\frac{I(Y_i - \boldsymbol{\beta}'\mathbf{X}_i \ge 0)}{\hat{G}(\boldsymbol{\beta}'\mathbf{X}_i)} - \frac{1}{2} \right] = 0 , \tag{2.7}$$

where $\hat{G}$ is the Kaplan-Meier estimate of the survival function of censoring distribution. Because of the discontinuity of the function $\mathbf{S}_n(\boldsymbol{\beta})$, the estimating equation (2.7) does not always have an exact solution, and therefore an estimator $\hat{\boldsymbol{\beta}}$ is defined as a minimizer of the Euclidean norm of the function $\| \boldsymbol{S}_n(\boldsymbol{\beta}) \|$.

Several other papers related to the median regression appeared lately in the literature. McKeague, Subramanian, and Sun (2001) introduced the median regression model of the same form as Ying et al. (1995), but used missing information principle to obtain the estimating equations for regression parameters under heavy censoring. Yin and Cai (2005) generalized the work by Ying et al. (1995) to the quantile regression for the correlated failure time data.

There appears to be only a few attempts to develop or describe a regression model for the residual life function. For parametric families of distributions possessing certain "setting the clock back to zero" property Rao, Damaraju, and Alhumoud (1993) illustrated the effect of the covariates on the percentile residual life function under the AFT assumption and proportionality of the hazard functions. In a Bayesian framework Gelfand and Kottas (2003) introduced the semiparametric median residual regression model which also was induced by the semiparametric accelerated failure time model.

Jeong et al. (2007) are currently working on time-specific median residual regression, where the median residual life function can be modeled at any time point specified a priori. More formally the regression model can be specified in the form

$$\log(\theta(t_0 \mid \mathbf{X}_i)) = \boldsymbol{\beta}'_{t_0} \mathbf{X}_i . \tag{2.8}$$

The authors propose to use a specific case of the estimating equation (2.7) appropriately modified for the time-specificity of the median residual model. This work can also be considered as a generalization of work by Ying et al. (1995).

3.0 PROPORTIONAL MEDIAN RESIDUAL LIFE MODEL

Literature review demonstrates that there is a gap in the inferential procedures for the median residual life function. Several regression methods available in the literature regress the MERL function on covariates at some specific time point (Ying, Jung and Wei, 1995; McKeague, Subramanian and Sun, 2001; Yin and Cai, 2005; Jeong, Jung and Bandos, 2007), are focused on a specific class of parametric distributions (Rao, Damaraju, and Alhumoud, 1993), or model the MERL function induced by the accelerated failure time assumption using the Bayesian approach (Gelfand and Kottas, 2003). We propose to fill this gap by introducing more general frequentist regression technique for the MERL function at multiple time points simultaneously. The proportional median residual life model has a conceptual similarity with the Cox proportional hazards model. The proposed regression technique can be used for modeling the proportionality of the MERL functions at multiple time points simultaneously over either the whole support interval or some pre-defined interval. We construct an estimating equation for parameter estimation and perform the simulation studies to assess the probability of type I error and power for testing the hypothesis of interest. We also apply the proposed method for analysis of a dataset from a breast cancer trial that was conducted by the NSABP.

3.1 PROPORTIONAL MEDIAN RESIDUAL MODEL

3.1.1 Model description and estimating equations

As before, let T_i and C_i denote failure and censoring time for the i^{th} patient respectively in a sample of size n, then $Y_i = \min(T_i, C_i)$ is the observed survival time and $\delta_i = I(T_i \le C_i)$ is the observed failure time indicator.

Let $\boldsymbol{X_i}$ be a p-dimensional covariate for T_i. We also assume that C_i is independent of T_i and $\boldsymbol{X_i}$, and $\{(T_i, C_i, \mathbf{X}_i), \quad i = 1,..,n\}$ are assumed to be independent and identically distributed. Also if we define $\theta(t \mid \boldsymbol{X}_i)$ as the median residual life function of T_i conditional on $\boldsymbol{X_i}$ we specify the form of the proportional median residual life model as

$$\theta(t \mid \boldsymbol{X_i}) = \theta_0(t)\exp(\boldsymbol{\beta' X_i}) \qquad (3.1),$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of covariates and $\theta_0(t)$ is an unspecified function which gives the median residual life function for a set of conditions $\boldsymbol{X_i} = \mathbf{0}$.

The proposed regression technique can be used for modeling the proportionality of the MERL functions at multiple time points simultaneously, regardless of whether they present the whole support interval or some pre-defined interval. Similarly to the Cox proportional hazards model, where the proportionality of the hazard functions regarded as constant over time, model (3.1) specifies proportionality of the median residual life functions to be constant over time. For the simplest case with one binary covariate, for example treatment group vs. control group, our proposed model specifies that median residual life functions of the control group and treated group are respectively $\theta_0(t)$ and $\eta\theta_0(t)$, where $\eta = e^{\beta}$ and β is the corresponding regression parameter. The positive value of the parameter estimate indicates an increase of the MERL

function for treated patients and value of $\eta = e^{\beta}$ implies the magnitude of the increase within an interval of interest. On the other hand a negative value of the parameter estimate indicates a decrease in the MERL function for treated patients and shows a negative effect of the intervention. Below is an example of two survival functions and their corresponding median residual life functions proportional over time with proportionality parameter 2. The data were generated from two exponential distributions with appropriately defined parameters.

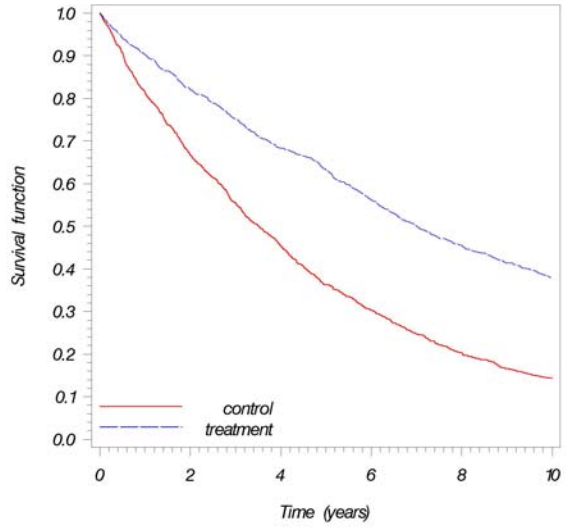

Figure 3-1 Survival functions with MERL functions proportional with parameter 2

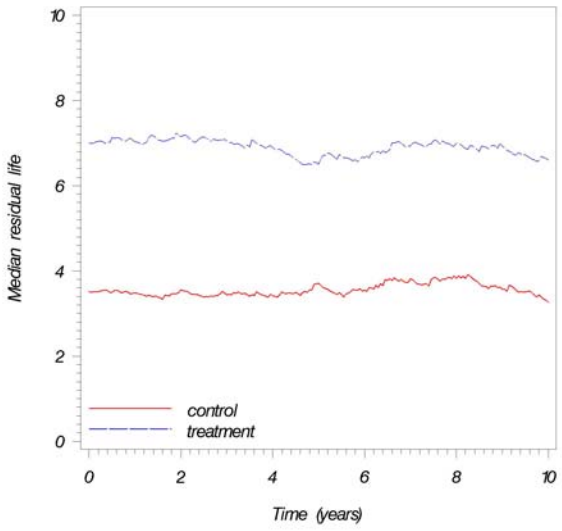

Figure 3-2 Median residual life functions proportional with parameter 2

As it was mentioned in section 2.2.3, a specific case of the estimating equation (2.7) was used in the work by Jeong et al. (2007) for estimation of the coefficients of the time-specific median residual regression (2.8). In the Appendix A, authors derive the estimating equation in the following form

$$\mathbf{S}_n(\boldsymbol{\beta}_{t_0}) = \sum_{i=1}^{n} \mathbf{X}_i \left[\frac{I(Y_i \geq t_0 + \exp(\boldsymbol{\beta}'_{t_0} \mathbf{X}_i))}{\hat{G}(t_0 + \boldsymbol{\beta}'_{t_0} \mathbf{X}_i)} - \frac{I(Y_i \geq t_0)}{2\hat{G}(t_0)} \right],$$

where $\hat{G}$ is the Kaplan-Meier estimate for the survival function of censoring distribution, or more formally is an estimator based on $\{(Y_i, 1-\delta_i), \quad i = 1,..,n\}$ set of pairs.

We extend the methodology used in the paper by Jeong et al. (2007) to construct the estimating equations for our model. Specifically, since in our new model the proportionality is assumed at every point of the interval of interest we consider averaging the estimating equation over that interval. For estimation of the regression parameters of the model (3.1) we introduce the function $\mathbf{S_n(\beta)}$ and propose to use it as an estimating function for $\boldsymbol{\beta}$ as follows

$$\mathbf{S}_n(\boldsymbol{\beta}) = \int_0^{\hat{T}} \sum_{i=1}^{n} \mathbf{X}_i \left\{ \frac{I(Y_i \geq t + \hat{\theta}_0(t)\eta_i)}{\hat{G}(t + \hat{\theta}_0(t)\eta_i)} - \frac{I(Y_i \geq t)}{2\hat{G}(t)} \right\} dt , \tag{3.2}$$

where $\eta_i = \exp(\boldsymbol{\beta}'\mathbf{X}_i), i = 1,..,n$, $\hat{\theta}_0(t)$ is a nonparametric estimator of the baseline MERL function, $\hat{G}$ is the Kaplan-Meier estimate for the survival function of censoring distribution and the interval of the integration $[0,\hat{T}]$ is determined from the data and will be discussed later. This function is a generalization of the estimating function used by Jeong et al. (2007).

3.1.2 Estimating procedure

An integral can be approximated as $\int_a^b f(t)dt \approx \sum_{j=1}^{m} f(t_j^*)\Delta_j$, where t_j^* is some arbitrary point in the interval Δ_j, and Δ_j is a partition of the interval $[a,b]$ such that $\max \Delta_j \to 0$. If we choose a partition such that all Δ_j are equal and t_j^* is a middle point of Δ_j, then our estimating function will have the form

$$\tilde{\mathbf{S}}_n(\boldsymbol{\beta}) = \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbf{X}_i \left\{ \frac{I(Y_i \geq t_j + \hat{\theta}_0(t_j)\eta_i)}{\hat{G}(t_j + \hat{\theta}_0(t_j)\eta_i)} - \frac{I(Y_i \geq t_j)}{2\hat{G}(t_j)} \right\}, \tag{3.3}$$

where t_j, $j = 1,..,m$, are the centers of the equal length intervals which partition some interval $[0,\hat{T}]$ chosen a priori. The rule for choosing time point $\hat{T}$ will be discussed in section 3.1.3. Δ_j's can be omitted from the definition of our estimating function as minimizing the sum $\sum_{j=1}^{m} f(t_j^*)\Delta_j$ is the same as minimizing the sum $\sum_{j=1}^{m} f(t_j^*)$ when all Δ_j's are equal.

Because of the discontinuity of $\tilde{\mathbf{S}}_\mathbf{n}(\boldsymbol{\beta})$, the estimating equation $\tilde{\mathbf{S}}_\mathbf{n}(\boldsymbol{\beta}) = 0$ does not always have an exact solution. In such situations it is a usual practice to minimize the Euclidean norm $\| \tilde{\mathbf{S}}_\mathbf{n}(\boldsymbol{\beta}) \|$, which leads to the approximate solution with the asymptotic behavior of the exact one (Vaart, 1998). We define an estimator $\hat{\boldsymbol{\beta}}$ as a minimizer of the Euclidean norm $\| \tilde{\mathbf{S}}_\mathbf{n}(\boldsymbol{\beta}) \|$, where the norm will be defined as the square root of sum of squares in our simulations and real example. Though Newton-Raphson optimization algorithm is a standard procedure for identifying an extremum of the function, it cannot be applied in our case because of the discontinuity of the function $\tilde{\mathbf{S}}_\mathbf{n}(\boldsymbol{\beta})$. We use a grid search method to minimize an integral approximation.

To obtain the required estimators we use an iterative procedure described below. The main idea of this procedure is to gradually increase the number of time points which are used to approximate an integral. We start from approximation at one point and continue to increase the number of points until some specified convergence criterion (*D*) is met. For brevity we use η_i instead of $\exp(\boldsymbol{\beta}'\mathbf{X}_i)$.

Step 1: Define $t_1 = T/2$ and using the grid search method obtain $\hat{\boldsymbol{\beta}}^{(1)}$ by minimizing the function

$$\tilde{\mathbf{S}}_n^{(1)}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X_i} \left\{ \frac{I(Y_i \geq t_1 + \hat{\theta}_0(t_1)\eta_i)}{\hat{G}(t_1 + \hat{\theta}_0(t_1)\eta_i)} - \frac{I(Y_i \geq t_1)}{2\hat{G}(t_1)} \right\}$$

Step 2: Define $t_1 = T/4, t_2 = 3T/4$ and using the grid search method obtain $\hat{\boldsymbol{\beta}}^{(2)}$ by minimizing the function

$$\tilde{\mathbf{S}}_n^{(2)}(\boldsymbol{\beta}) = \sum_{j=1}^{2}\sum_{i=1}^{n}\mathbf{X}_i\left\{\frac{I(Y_i \geq t_j + \hat{\theta}_0(t_j)\eta_i)}{\hat{G}(t_j + \hat{\theta}_0(t_j)\eta_i)} - \frac{I(Y_i \geq t_j)}{2\hat{G}(t_j)}\right\}$$

Obtain the "distance" between $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$, which can be defined as Euclidean norm $d_1 = \| \hat{\boldsymbol{\beta}}^{(1)} - \hat{\boldsymbol{\beta}}^{(2)} \|$ and compare d_1 to the prespecified convergence criterion constant D. If the distance d_1 is less then D, report $\hat{\boldsymbol{\beta}}^{(2)}$ as a solution of the estimating equation $\tilde{\mathbf{S}}_\mathbf{n}(\boldsymbol{\beta}) = 0$ otherwise continue to the next iteration step.

Step 3: Define $t_1 = T/6, t_2 = T/2, t_3 = 5T/6$ and using the grid search method obtain $\hat{\boldsymbol{\beta}}^{(3)}$ by minimizing the function

$$\tilde{\mathbf{S}}_n^{(3)}(\boldsymbol{\beta}) = \sum_{j=1}^{3}\sum_{i=1}^{n}\mathbf{X}_i\left\{\frac{I(Y_i \geq t_j + \hat{\theta}_0(t_j)\eta_i)}{\hat{G}(t_j + \hat{\theta}_0(t_j)\eta_i)} - \frac{I(Y_i \geq t_j)}{2\hat{G}(t_j)}\right\}$$

Obtain the "distance" between $\hat{\boldsymbol{\beta}}^{(2)}$ and $\hat{\boldsymbol{\beta}}^{(3)}$, $d_2 = \| \hat{\boldsymbol{\beta}}^{(2)} - \hat{\boldsymbol{\beta}}^{(3)} \|$ and compare d_2 to prespecified convergence criterion constant D. If the convergence criterion is met, report $\hat{\boldsymbol{\beta}}^{(3)}$ as a solution of the estimating equation $\tilde{\mathbf{S}}_\mathbf{n}(\boldsymbol{\beta}) = 0$ otherwise continue to the next step of the iteration process. Continue this procedure until the prescribed convergence criteria are met. Finally we obtain the sequence of the parameter estimates $\hat{\boldsymbol{\beta}}^{(m)}, m = 1,..,M$ and we define our final parameter estimate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(M)}$.

At each iteration step k the number of points used to approximate the integral equals k.

3.1.3 Estimation interval

Estimation interval where proportionality is assumed can be chosen using two approaches. The proportionality of the MERL functions can be assumed on some predefined interval of the entire follow-up period. This choice of the interval can be based on the data or personal believes of an investigator. Also the proportionality of the MERL functions can be assumed on the whole interval where the estimator of the MERL function is properly defined. If the interval is chosen using this method, some difficulties could be experienced. As time progresses the estimates of the survival function become less reliable and more unstable since the number of events gradually decrease. Therefore in practice the estimates of the median residual life function at time points close to the largest time on study might be unreliable, even though the MERL function is still formally defined. Because of the certain arbitrariness of choosing the range of integration for equation (3.2) we can attempt to improve the efficiency of the estimation by considering an interval of integration that is smaller than the interval where the estimator of the MERL function is properly defined. In this work we use the following formula to define time point $\hat{T}$ such that $[0,\hat{T}]=[0,\hat{S}_0^{-1}(2\hat{S}_0(t_\bullet))]$, where $t_\bullet$ is the event before last one for the baseline group and $\hat{S}_0(t)$ is the Kaplan-Meier estimate of the survival function for the baseline group

3.1.4 Variance of the parameter estimates

Statistical inferences about the regression parameter can be simplified by availability of the variance of the parameter estimate. However in our case the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ depends on the distribution of the error terms which cannot be easily estimated. We propose to

use resampling techniques for variance estimation and use bootstrap method (Efron, 1981) in our simulations and real-data example. More specifically we draw a simple random sample $\{(Y_i^*, \delta_i^*, \mathbf{X}_i^*), \quad i = 1,..,n\}$ with replacement from the original data $\{(Y_i, \delta_i, \mathbf{X}_i), \quad i = 1,..,n\}$ with equal probability *1/n* and for each bootstrap realization we estimate $\hat{\boldsymbol{\beta}}_j^*$. After this procedure is performed *B* times we estimate variance-covariance matrix based on the bootstrap sample of the parameter estimates $\{\hat{\boldsymbol{\beta}}_1^*,..,\hat{\boldsymbol{\beta}}_B^*\}$. The estimate of the standard error of the regression parameter can later be used to perform a Wad type test on the parameters.

3.1.5 Checking the proportionality assumption

One of the approaches for checking the proportionality assumption in two groups, which is the simplest case of the regression $\theta(t) = \theta_0(t)\exp(\beta_1 X_1)$, is a graphical one. For a given data one can plot the natural logarithm of the nonparametric estimates of the median residual life function in one group vs. the other group. From the functional form of the proposed model the following will be true

$$\log(\theta(t)) = \log(\theta_0(t)) + \beta_1 .$$

Therefore if the proportionality assumption holds on some prespecified interval, the graph of $\log(\theta(t))$ vs. $\log(\theta_0(t))$ would resemble a straight line with an intercept close to β_1.

This graphical check is simple to perform as MERL function can be easily estimated using the Kaplan-Meier estimator of the corresponding survival curves and the equation (2.3).

3.1.6 Parametric distributions and proportional MERL model

Since the MERL function does not uniquely define the survival distribution, the problem of estimating the parameters of the model arises even for the parametric approach. Namely, if we assume that the baseline distribution belongs to a certain parametric family, it is not clear whether the distribution filtered through the proportional MERL model belongs to the same family. Below we show that exponential distribution guarantees a one-to-one correspondence between the survival function and median residual life function under the assumption of proportionality of the MERL functions.

Let's assume that an exponential distribution defines the baseline distribution $T_0 \sim EXP(\lambda)$ and the proportional MERL model (3.1) is satisfied. The survival function for the baseline is then defined as $S_0(t) = e^{-\lambda t}$ and its inverse can be calculated as $S_0^{-1}(y) = -\frac{1}{\lambda}\ln(y)$. Therefore, using (2.1), the MERL function corresponding to the baseline is given by $\theta_0(t) = \frac{1}{\lambda}\ln 2$. We search for a distribution for the variable T with the MERL function $\theta(t)$ proportional to the baseline with the factor η, $\theta(t) = \eta\theta_0(t)$, within the exponential family. Therefore $\theta(t) = \frac{\eta}{\lambda}\ln 2$ and it uniquely defines a distribution within the exponential family with parameter λ/η, $T \sim EXP(\frac{\lambda}{\eta})$.

We use this fact to perform our simulation studies.

3.2 SIMULATION STUDY

3.2.1 Simulation scenarios

We performed numerical studies to investigate the finite sample properties of the proposed inference procedure based on the estimating function (3.3). Simple proportional median residual life model was assumed which included one binary covariate and took a form

$$\theta(t) = \theta_0(t) exp(\beta_1 X_1) . \tag{3.5}$$

Covariate X_1 was generated from a Bernoulli distribution with probability of success 0.5. Three scenarios were considered for the censoring proportion – 0%, 10% and 20% censoring. Failure times were simulated from an exponential distribution and we set the rate parameter for the baseline to be $\lambda = 0.2$. For each numerical study we simulated n observations from the exponential distribution with parameter $\frac{\lambda}{\exp(\beta_1 X_{1i})}$, $i = 1,..,n$. To generate failure times from the corresponding distribution we used the probability integral transformation technique (Casella and Berger, 2002). First, n observations were generated from a uniform distribution over the interval (0, 1) and then the inverse of the exponential distribution transformation was applied as follows,

$$T_i = -\frac{\exp(\beta_1 X_{1i})}{\lambda} \ln(u_i) \quad i = 1,..,n ,$$

where u_i is from the uniform distribution between 0 and 1. The censoring times C_i's were generated from the uniform distribution between 0 and c, where c is a constant that controls for the censoring proportion. Then the observed data were determined by $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \le C_i)$.

The grid search algorithm was used to minimize the estimating function (3.3). In practice if $\hat{G}(t_j + \hat{\theta}_0(t_j)\eta_i)$ and $\hat{G}(t_j)$ are zeros in (3.3) then $\frac{I(Y_i \geq t_j + \hat{\theta}_0(t_j)\eta_i)}{\hat{G}(t_j + \hat{\theta}_0(t_j)\eta_i)}$ and $\frac{I(Y_i \geq t_j)}{2\hat{G}(t_j)}$ are also set to be zeroes correspondingly. Also to decrease the amount of time required for these extensive simulations we did not use the iteration procedure described in section 3.1.2, but instead we fixed the number of time points required for the integral estimation. We used four accordingly chosen time points on interval $[0, T]$. The interval of approximation was chosen for each simulated dataset according the rule described in section 3.1.3 using the equation (3.4) as $T = \hat{S}_0^{-1}[2\hat{S}_0(t)]$, where time point *t* was defined as the time of the event before the last one in the baseline group.

3.2.2 Simulation results

For the purpose of estimating the bias and standard deviation of the parameter estimates 1000 simulations were generated for each configuration of sample sizes of 50, 100, 150 and 200 and censoring percent of 0, 10 and 20. These results are presented in *Table 3-1*. For each data realization of sample size of 200 we draw 400 bootstrap samples to estimate the standard error of the regression parameter. In *Table 3-2* we present the sample standard deviation of the 1000 estimates (SD), the square root of the average bootstrap variances based on 400 bootstrap samples for each data realization (SE_b), and the average length of the estimation interval (T_{end}). Probabilities of type I error were calculated based on the Wald test statistic. For the purpose of estimating these probabilities the data with the sample size of 200 were generated with the true β_1 being equal to 0. Again 1000 simulations for each scenario were used for this purpose. To

investigate the power of the Wald's test, the data were generated with β_1 equal to 0.5 and 0.7. 500 simulations for each choice of the censoring proportion with sample size of 200 were used for the estimation of power.

Table 3-3 presents the probabilities of rejecting the null hypothesis $H_0 : \beta_1 = 0$ when the true regression parameter equals 0, 0.5 and 0.7 respectively. These probabilities were based on the Wald statistic and reflect the probability of type I error, when true β = 0 and power, when β = 0.5 and β = 0.7.

Table 3-1 Proportional MERL model (empirical bias and standard deviation)

	Average censoring proportion					
	0%		10%		20%	
n	$\Delta\beta_1$	SD	$\Delta\beta_1$	SD	$\Delta\beta_1$	SD
50	-0.011	0.403	0.004	0.450	0.015	0.488
100	-0.025	0.292	0.003	0.318	-0.003	0.329
150	0.010	0.246	-0.017	0.255	0.008	0.262
200	-0.001	0.215	-0.012	0.217	0.010	0.232

Table 3-2 Proportional MERL model (bootstrap standard error and average length of the estimation interval, n = 200)

c%	$\Delta\beta_1$	SD	SE_b	T_{end}
0	-0.001	0.215	0.238	18.43
10	-0.012	0.217	0.242	15.30
20	0.010	0.232	0.252	11.51

Table 3-3 Proportional MERL model (the rejection rate when the true regression parameter is β, n = 200)

c%	β = 0	β = 0.5	β = 0.7
0	0.029	0.646	0.864
10	0.034	0.556	0.834
20	0.034	0.520	0.802

From *Table 3-1* it can be seen that the parameter estimates are approximately unbiased. As it is expected, the standard deviations of the parameter estimates across 1000 simulations increase

with higher censoring proportion and decrease with larger sample size. From *Table 3-2* the bootstrap standard errors which are summarized by the square root of the average of the bootstrap variances seem to provide fair estimates of the variability compared to the standard deviations. As it is seen from the table, the bootstrap standard errors slightly overestimate the variance, but they still reflect a stable pattern of increased variability as the censoring proportion increases. Also as it was expected, the width of the interval of estimation decreases as the censoring proportion increases.

Table 3-3 agrees with our previous observations that the bootstrap resampling technique overestimates the standard errors of the corresponding parameter estimates resulting in conservative conclusions. Power decreases with higher censoring proportion and increase when the true value of the regression parameter moves away from the null value.

3.3 EXAMPLE

For the illustration purpose we apply the proposed method to the NSABP protocol B-04 dataset (Fisher et al., 2002). This dataset is a typical example with a long-term follow-up, as it contains survival information among breast cancer patients for over 30 years. The total number of eligible patients accrued for this trial was 1665 and the censoring proportion was about 23 percent. In this trial there were 5 groups being compared – three groups in node-negative patients, and two groups in node-positive patients. The purpose of this study was to compare the effects of total mastectomy and radical mastectomy with or without postoperative radiation therapy on overall survival. In this dissertation we use the nodal status by itself as one of the covariates. We also use the pathological tumor size as another covariate of interest, which originally is a continuous

covariate, but here is categorized at the median into two groups – those patients with this characteristic below its 50^{th} percentile and those above its 50^{th} percentile.

We fit two univariate models – Model 1 with the node status as a single covariate and Model 2 with the categorized pathological tumor size as a single covariate (we code it as 0 for those patients with the tumor size ≤ 3cm and 1 for those patients with the tumor size > 3cm), and one multivariate model, Model 3, which incorporates both of the prognostic factors. 66 patients were deleted from the dataset for the analysis of the data to use models 2 and 3 because of the unknown tumor size characteristic. Based on some preliminary analysis, we assume the proportionality of the median residual life functions on interval [0, 5] for Model 1, on interval [0, 2] for Model 2, and the assumed interval of proportionality for Model 3 is taken as [0, 2] which is the smallest interval of the previous two. We use a graphical way of assessing model performance by plotting the nonparametric and model-based estimates of the MERL function on the same graph.

To obtain the parameter estimates we used the iteration scheme described in the section 3.1.2, where the number of points required for the integral approximation is increased by one at each following step. We continued this procedure until the convergence level of 0.01 is satisfied. To estimate the standard errors of the regression parameters, the bootstrap resampling technique was used with 1000 bootstrap samples taken for each model. Due to the substantial amount of time required for the parameter estimation for each bootstrap sample, we fixed the number of points required for the iteration process to converge in the original data and used it for each bootstrap sample. The number of points required for the procedure to converge up to the specified convergence level of 0.01 was three for models 1 and 2 and two for model 3.

In Table 3-4 we present the corresponding parameter estimates ($\hat{\beta}$), their standard errors, calculated using the bootstrap resampling method (SE_b), the Wald test statistic $z = \hat{\beta} / \widehat{SE}(\hat{\beta})$ and the corresponding *p*-value.

Table 3-4 B-04 results for the proportional MERL model

Model	Variable	$\hat{\beta}$	SE_b	z	p-value
Model 1	Node	-0.582	0.0915	- 6.366	< 0.0001
Model 2	Paths	-0.381	0.0968	- 3.932	0.0001
Model 3	Node	-0.540	0.0970	-5.570	< 0.0001
	Paths	-0.235	0.0758	-3.096	0.0020

All three models show high statistical significance of the variables in the model. In model 1 the parameter estimate for the effect of node status was -0.582, which indicated a decrease in median residual life for the node positive patients by approximately 44% compared to the node negative patients over the first five years $(1 - e^{-0.582})$. In model 2 the estimate of the regression coefficient corresponding to tumor size was -0.381, which also indicated a decrease in MERL for the patients with pathological tumor size > 3cm by approximately 32% compared to the patients with the tumor size of 3 cm or less over the first two years $(1 - e^{-0.381})$. The joint effect of these two

variables, estimated from model 3, decreases the MERL of the patients over the first two years by 54% compared to the patients in the baseline group $(1-e^{-0.540-0.235})$. *Figure 3-3*, *Figure 3-4* and *Figure 3-5* show the nonparametric estimates of the median residual life functions for each subgroup defined by the covariates and the estimates of the MERL functions evaluated by the parameter estimates from the corresponding model. We can see from the graphs that nonparametric estimates of the MERL functions are very close to the model-based MERL functions.

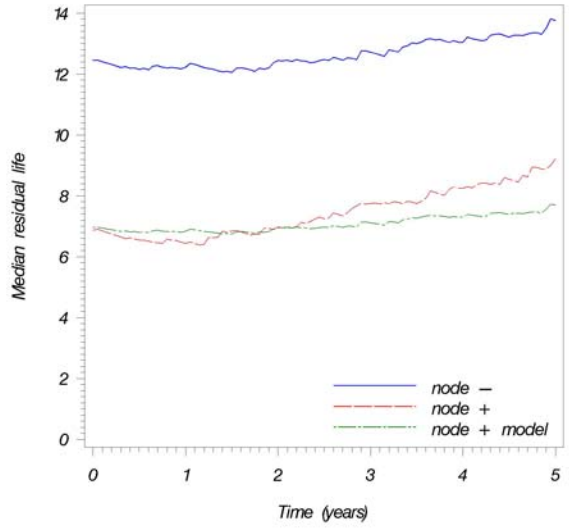

Figure 3-3 B-04 node status as a covariate

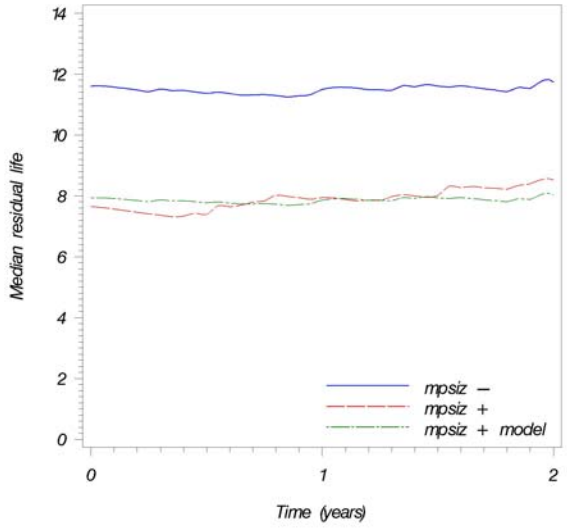

Figure 3-4 B-04 pathological tumor size as a covariate

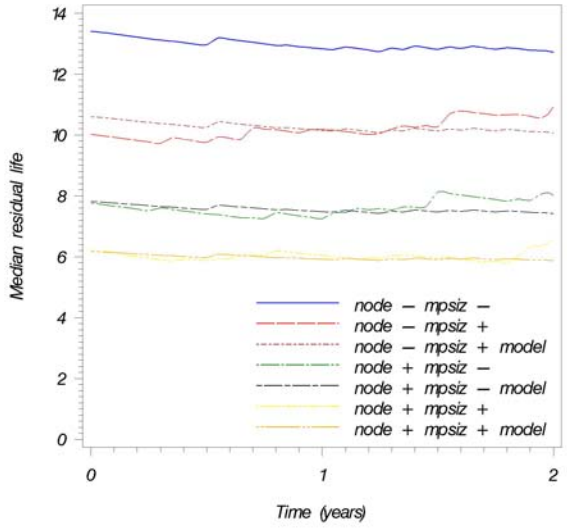

Figure 3-5 B-04 node status and pathological tumor size as covariates

3.4 DISCUSSION

In this chapter of the dissertation we have defined and developed the proportional median residual life model. The structure of the model shares a certain similarity with the Cox proportional hazards model, namely it assumes the constant proportionality of MERL functions over the time interval of interest. The estimates of the regression parameters are obtained using iterative solution to the estimating equations, and their corresponding standard errors are computed using the bootstrap resampling technique.

This regression presents a novel approach to model the relationship between the median residual life function and the covariates of interest at multiple time points simultaneously. Such model may be of a significant importance to clinicians and medical researchers as the concept of the median residual life function is clinically relevant and intuitively appealing without additional statistical details. Also the model can be used to compare two or more groups of interest, such as treatment groups, by means of median residual life function adjusting for the important covariates, such as age, gender, blood pressure and so on.

One of the additional advantages of the proposed regression method and corresponding estimation technique is that they provide the researcher with a substantial flexibility in assumption of proportionality. The method allows for choosing the interval of estimation based on the data and personal believes of the investigator. If someone is willing to assume the proportionality of the median residual life functions only for a subset of the entire follow-up period, our method allows for doing so without losing any data. On the contrary if the same has to be done for the Cox proportional hazards model, i.e. assume the proportionality of the hazard functions only until a certain time point t, all subjects that have experienced an event after the

time point t would have to be censored at time t, which by reducing the number of events could substantially increase the censoring proportion.

Although our new regression model has a number of good properties, it has some limitations. As the median residual life functions may converge to each other as time passes, it would probably be unrealistic to assume the constant proportionality over time, especially for overall survival as an event of interest. Though this may be a problem for the entire study period, our model allows for assuming the proportionality on fixed interval and estimating the parameters of interest on that interval without any loss of the data.

Among the list of known and widely used distributions, only exponential distribution was identified as one that possess the property of one-to-one correspondence between the MERL and survival function under the proportional MERL model.

An estimating equation (3.3) also has its own disadvantages, as it requires the capability to estimate the median residual life function for the baseline group. The higher the number of groups defined by all combinations of the variables in the model, the more difficult this task becomes as the categorization decreases the number of observations in each subgroup which makes the estimate of the baseline median residual life function less reliable. Another difficulty arises when one of the covariates of interest is continuous. In this case some categorization of this covariate has to be done a priori. The idea to discretize a covariate into K groups was proposed by Ying, Jung and Wei, (1995) to fit the median regression when the assumption of independence between the censoring distribution and the vector of covariates is not satisfied.

Quantile statistics in general and the median residual life function in particular cannot be reliably estimated unless the censoring proportion is below some level. For example, the simple median is easily and reliably estimated if the censoring proportion is below 50%. This fact leads

to some limitations on the type of data that can be used for the proposed regression models. In some instances special techniques can be applied to account for the high censoring proportion, such as missing information principle that was used in McKeague, Subramanian and Sun (2001).

One of the major disadvantages of the proposed method is time required for the estimation of the parameters and especially their corresponding standard errors. The bootstrap resampling technique in general is a very computationally intensive method. Also the amount of required time increases substantially as the dimension of the vector of regression coefficients increases.

Determining the minimum of the function is another computational difficulty that arises in the process of estimating of the regression coefficients. The grid search method is the technique applied in this work. While it is one of the elementary yet robust techniques for the required task, it is also a very computationally intensive method as the amount of time required for the convergence increases substantially as the dimension of the vector of parameters increases. Also it may not converge to the global extremum of the function for some instances.

The problem of defining the number of points required for the integral approximation also requires a special attention. In our present work each new iteration step increases the number of points in the integral approximation by one. Since every iteration step is followed by minimization of the function of interest, such a slow increase in the number of points required for the integral approximation might slow down the overall convergence of the algorithm. On the other hand, a more aggressive increase of number of points (e.g. by more than one) could still lead to an unnecessary computer intensive iteration step also slowing down the overall convergence.

4.0 ACCELERATED MEDIAN RESIDUAL LIFE MODEL

In this chapter, another type of regression on the median residual life function is proposed, which by its analytical form resembles the accelerated failure time model. Parametric approach for the model fitting is discussed and numerical studies are performed to investigate the empirical bias of the parameter estimates, the probability of type I error and power of the proposed statistical test under different scenarios. The relation between the proposed model and the accelerated failure time model is presented. For the illustration purposes a dataset is simulated from a Weibull distribution and two methods of estimation are compared.

4.1 ACCELERATED MERL MODEL

As before, let T_i defines failure time for the i^{th} patient in a sample of size n and C_i defines censoring time for the patient i, then $Y_i = \min(T_i, C_i)$ is the observed time and $\delta_i = I(T_i \le C_i)$ is the observed failure time indicator. Let $\boldsymbol{X_i}$ be a p-dimensional covariate for T_i. We also assume that C_i is independent of T_i and $\boldsymbol{X_i}$, and $\{(T_i, C_i, \mathbf{X}_i), \quad i = 1,..,n\}$ are independent and identically distributed. Also if we define $\theta(t \mid \boldsymbol{X}_i)$ as the median residual life function of T_i, a conditional on $\boldsymbol{X_i}$, we specify the form of the accelerated median residual life model as

$$\theta(t \mid \boldsymbol{X}_i) = \exp(\boldsymbol{\beta' X}_i)\theta_0(t\exp(-\boldsymbol{\beta' X}_i)), \tag{4.1}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of covariates and $\theta_0(t)$ is an unspecified function which gives the median residual life function for a set of conditions $\boldsymbol{X_i} = \mathbf{0}$.

The analytical from of this model is similar to the accelerated failure time model. For the simplest case of the regression with binary predictor, for example treatment group vs. control group, model (4.1) indicates that the median residual life functions of the control group and treated group are respectively $\theta_0(t)$ and $\eta\theta_0(t/\eta)$, where $\eta = e^{\beta}$ and β is the corresponding regression parameter. The positive estimate of the regression coefficient indicates an increase of the MERL functions for treated patients with a shift in the time axis. Below is an example of two survival functions and their corresponding median residual life functions under the accelerated MERL model with $\eta = 2$. The data were generated from a Weibull distribution with appropriately defined parameters.

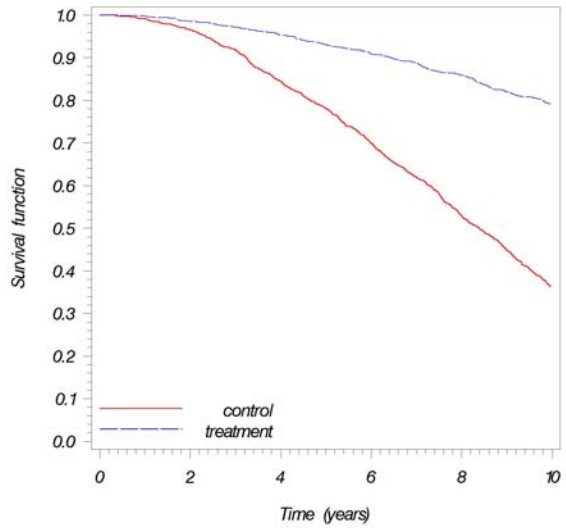

Figure 4-1 Survival functions with MERL functions accelerated by the factor 2

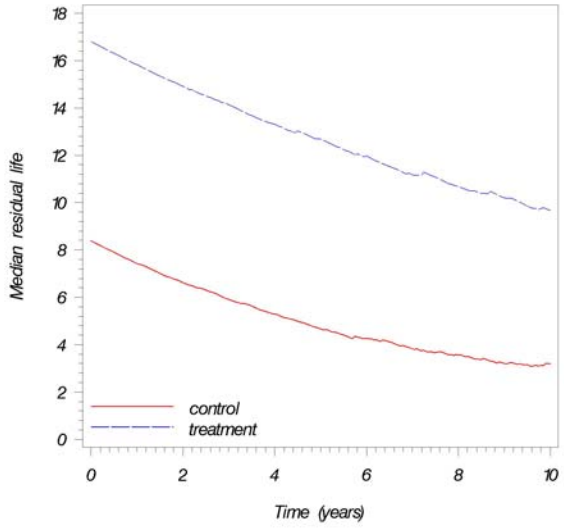

Figure 4-2 Median residual life functions accelerated by the factor 2

4.2 PARAMETRIC APPROACH

4.2.1 Parametric distributions and accelerated MERL model

Since the MERL function does not uniquely define the survival distribution, the problem of estimating the parameters of the model arises even for the parametric approach. Namely, if we assume that the baseline distribution belongs to a certain parametric family, it is not clear, in general, whether a distribution filtered through the accelerated MERL model belongs to the same family. However, some of the distribution families may guarantee one-to-one relationship between the median residual life function and survival function under the accelerated median residual life model. Restricting modeling to such families allows for avoiding the problem of

nonuniqueness. Below we demonstrate that Weibull, exponential power and Jeong (2006) distributions are among such families.

a) Accelerated MERL model within Weibull distribution.

Let's assume that a Weibull distribution $WEI(\lambda,\kappa)$ defines the baseline distribution and model (4.1) is satisfied. The survival function for the baseline is then defined as $S_0(t)=e^{-(\lambda t)^{\kappa}}$ and its inverse can be calculated as $S_0^{-1}(y)=\frac{1}{\lambda}(-\ln(y))^{1/k}$. Therefore, using (2.1), the MERL function corresponding to the baseline is calculated as $\theta_0(t)=\frac{1}{\lambda}(\ln 2+(\lambda t)^k)^{1/k}-t$. We search for a distribution for the variable T with the MERL function $\theta(t)$ accelerated by the factor η, $\theta(t)=\eta\theta_0(\frac{t}{\eta})$, within the Weibull family. Therefore $\theta(t)=\frac{\eta}{\lambda}(\ln 2+(\frac{\lambda}{\eta}t)^k)^{1/k}-t$ and it uniquely defines a distribution within the Weibull family with parameters λ/η and κ respectively.

$$\left.\begin{array}{l} T_0\sim WEI(\lambda,\kappa) \\ \theta(t)=\eta\theta_0(\frac{t}{\eta}) \\ T\sim WEI(\lambda_1,\kappa_1) \end{array}\right\}\Rightarrow\left\{\lambda_1=\frac{\lambda}{\eta}\text{ and }\kappa_1=\kappa\right\}\Rightarrow T\sim WEI(\frac{\lambda}{\eta},\kappa)$$

b) Accelerated MERL model within exponential power distribution

Let's assume that an exponential power distribution $EP(\lambda,\kappa)$ defines the baseline distribution and model (4.1) is satisfied. The survival function for the baseline is then defined as $S_0(t)=\exp(1-e^{(\lambda t)^{\kappa}})$ and its inverse can be calculated as $S_0^{-1}(y)=\frac{1}{\lambda}(\ln(1-\ln(y)))^{1/k}$. Therefore, using (2.1), the MERL function corresponding to the baseline is calculated as $\theta_0(t)=\frac{1}{\lambda}(\ln(\ln 2+e^{(\lambda t)^k}))^{1/k}-t$. We search for a distribution for the variable T with the MERL

function $\theta(t)$ accelerated by the factor η, $\theta(t)=\eta\theta_0(\frac{t}{\eta})$, within the exponential power family.

Therefore $\theta(t)=\frac{\eta}{\lambda}(\ln(\ln 2+e^{(\frac{\lambda}{\eta}t)^k}))^{1/k}-t$ and it uniquely defines a distribution within the exponential power family with parameters λ/η and κ respectively.

$$\left.\begin{array}{l} T_0 \sim EP(\lambda,\kappa) \\ \theta(t)=\eta\theta_0(\frac{t}{\eta}) \\ T \sim EP(\lambda_1,\kappa_1) \end{array}\right\} \Rightarrow \left\{\lambda_1=\frac{\lambda}{\eta} \text{ and } \kappa_1=\kappa\right\} \Rightarrow T \sim EP(\frac{\lambda}{\eta},\kappa)$$

c) Accelerated MERL model within Jeong distribution

Let's assume that a Jeong distribution $JEO(\alpha,\kappa,\rho,\tau)$ defines the baseline distribution and model (4.1) is satisfied. The survival function and its corresponding inverse for the baseline are then defined as $S_0(t)=\exp\left\{-\frac{\alpha^{1-\tau}\{(\rho t)^\kappa+\alpha\}^\tau-\alpha}{\tau}\right\}$ and $S_0^{-1}(y)=\frac{1}{\rho}\left[\left\{\frac{\alpha-\tau\ln y}{\alpha^{1-\tau}}\right\}^{1/\tau}-\alpha\right]^{1/\kappa}$.

Therefore, using (2.1), the MERL function corresponding to the baseline is calculated as $\theta_0(t)=\frac{1}{\rho}\left[(\alpha^{\tau-1}\tau\ln 2+\{(\rho t)^\kappa+\alpha\})^{1/\tau}-\alpha\right]^{1/\kappa}-t$. We search for a distribution for the variable T with the MERL function $\theta(t)$ accelerated by the factor η, $\theta(t)=\eta\theta_0(\frac{t}{\eta})$, within the Jeong family. Therefore $\theta(t)=\frac{\eta}{\rho}\left[(\alpha^{\tau-1}\tau\ln 2+\{(\frac{\rho}{\eta}t)^\kappa+\alpha\})^{1/\tau}-\alpha\right]^{1/\kappa}-t$ and it uniquely defines a distribution within the Jeong family with parameters α, κ, ρ/η and τ respectively.

$$\left.\begin{array}{l} T_0 \sim JEO(\alpha,\kappa,\rho,\tau) \\ \theta(t)=\eta\theta_0(\frac{t}{\eta}) \\ T \sim JEO(\alpha_1,\kappa_1,\rho_1,\tau_1) \end{array}\right\} \Rightarrow \left\{\alpha_1=\alpha,\kappa_1=\kappa,\rho_1=\frac{\rho}{\eta} \text{ and } \tau_1=\tau\right\} \Rightarrow T \sim JEO(\alpha,\kappa,\frac{\rho}{\eta},\tau)$$

Restriction to a specific class of distributions does not always allow for circumventing the non-uniqueness problem. Pareto distribution is an example of such distribution where survival function is not uniquely determined even within the class of Pareto distributions. Let's assume that Pareto distribution $PAR(\lambda,\kappa)$ defines the baseline distribution and model (4.1) is satisfied. The survival function for the Pareto distribution is defined as $S_0(t)=\left(\frac{\lambda}{t}\right)^{\kappa}$ and the inverse of this function can be written as $S_0^{-1}(y)=\lambda y^{-1/\kappa}$. Therefore the median residual life function of the baseline distribution equals $\theta_0(t)=(2^{1/\kappa}-1)t$. If we assume that the accelerated median residual life model $\theta(t)=\eta\theta_0(\frac{t}{\eta})$ is correct then $\theta(t)=(2^{1/\kappa}-1)t$, which corresponds to the whole family of Pareto distributions with parameter *k* and any shape parameter λ_1.

$$\left.\begin{array}{l} T_0 \sim PAR(\lambda,\kappa) \\ \theta(t)=\eta\theta_0(\frac{t}{\eta}) \\ T \sim PAR(\lambda_1,\kappa_1) \end{array}\right\} \Rightarrow \left\{\lambda_1 \in \mathbb{R}^+ \text{ and } \kappa_1=\kappa\right\} \Rightarrow T \sim PAR(\lambda_1,\kappa) \quad \forall \lambda_1 > 0$$

One-to-one correspondence between the MERL and survival function under the accelerated median residual life model for some well-known and widely-used distributions can be used to fit this regression under the parametric setting. The maximum likelihood technique can be easily implemented for the estimation of the regression coefficients and their standard errors. The properties of the ML estimators can also be used to perform the hypothesis testing for the parameters of interest.

4.2.2 Maximum likelihood estimation

Assumption of a specific parametric form for the distribution for failure time *T* allows for using the maximum likelihood (ML) approach for inferential purposes. The invariance property of the maximum likelihood estimators can be used to calculate the ML estimator of the median residual life function for the cases when survival function is available in the closed form. For example, for the Weibull distribution the ML estimator of the corresponding MERL function equals $\hat{\theta}(t) = \frac{1}{\hat{\lambda}}(\ln 2 + (\hat{\lambda} t)^{\hat{k}})^{1/\hat{k}} - t$, where $(\hat{\lambda}, \hat{\kappa})$ are the ML estimators of the parameters λ and κ. The delta method provides a way to approximate the variance of the function of the MLEs for a large sample and therefore we can estimate the variance of the MERL function as a function of time *t*. In general, if ϕ defines a vector of parameters, $\hat{\phi}$ defines a ML estimator of the vector ϕ and $\theta(\phi)$ defines the function of interest, then the variance of the function $\hat{\theta}(\phi)$ can be asymptotically calculated as follows

$$Var\{\hat{\theta}(\phi)\} = \left(\frac{\partial \theta}{\partial \phi}\right)'_{\phi=\hat{\phi}} Var(\hat{\phi}) \left(\frac{\partial \theta}{\partial \phi}\right)_{\phi=\hat{\phi}},$$

where $Var(\hat{\phi})$ is the variance-covariance matrix of the ϕ, $\partial\theta / \partial\phi$ is the column vector of the first derivatives of the function θ with respect to the parameter vector ϕ and "′"denotes the transposed vector.

To be able to use the maximum likelihood approach, the likelihood function has to be defined. In general, if T_i is failure time for the i^{th} patient in the sample of size *n*, C_i is censoring time for a patient *i*, $Y_i = \min(T_i, C_i)$ is the observed time and $\delta_i = I(T_i \le C_i)$ is the observed failure time indicator, and we assume, that $T_1, ..., T_n \sim f(t), S(t)$, where $f(t)$ defines the

probability density function and $S(t)$ defines the corresponding survival function, then the loglikelihood function can be written as $Log(L) = Log(\prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i})$. After this function has been defined, the standard maximization procedures can be used to obtain the maximum likelihood estimators of the parameters and their standard errors. When the maximum of the loglikelihood function is not available in the closed form, some numerical methods are applied. The Newton-Raphson method is the most commonly used technique to obtain the extremum of the function of interest.

4.2.3 MLE for the Weibull distribution

In the example of the Weibull distribution we demonstrate the general technique of defining the likelihood function and using it for the parameter estimation. We assume that both location and scale parameters of the baseline Weibull distribution are unknown. To set up the likelihood function the probability density function and corresponding survival function of the distribution of interest have to be available.

We assume that $T_i \sim WEI(\lambda \exp(-\boldsymbol{\beta}'\mathbf{X}_i), \kappa), \quad i = 1,..,n$ and define $R = \sum_{i=1}^{n} \delta_i$ as the number of events in the sample, then

$$f_i(t) = \kappa\{\lambda \exp(-\boldsymbol{\beta}'\mathbf{X}_i)\}^k t^{\kappa-1} e^{-(\lambda \exp(-\boldsymbol{\beta}'\mathbf{X}_i)t)^k}, \quad i = 1,..,n$$

$$S_i(t) = e^{-(\lambda \exp(-\boldsymbol{\beta}'\mathbf{X}_i)t)^k}, \quad i = 1,..,n$$

$$Log(L) = Log(\prod_{i=1}^{n} \kappa^{\delta_i} \{\lambda \exp(-\boldsymbol{\beta}'\mathbf{X}_i)\}^{k\delta_i} y_i^{(\kappa-1)\delta_i} e^{-(\lambda \exp(-\boldsymbol{\beta}'\mathbf{X}_i) y_i)^k}) =$$

$$= Log(k^R \lambda^{kR} \prod_{i=1}^{n} \exp(-\kappa\delta_i \boldsymbol{\beta}'\mathbf{X}_i) y_i^{(\kappa-1)\delta_i} e^{-(\lambda^\kappa \exp(-\kappa\boldsymbol{\beta}'\mathbf{X}_i) y_i^\kappa)}) =$$

$$= R\ln(\kappa) + \kappa R \ln(\lambda) - \kappa \sum_{i=1}^{n} \delta_i \boldsymbol{\beta}'\mathbf{X}_i + (\kappa-1)\sum_{i=1}^{n} \delta_i \ln(y_i) - \lambda^\kappa \sum_{i=1}^{n} \exp(-\kappa\boldsymbol{\beta}'\mathbf{X}_i) y_i^\kappa.$$

As no closed form solutions exist for the parameter estimates and their standard errors in case of the Weibull distribution, we use the Newton-Raphson method to calculate these estimates.

4.2.4 Simulation study

For the simulation purposes we generate the data under condition of the accelerated median residual life model, where the baseline distribution is assumed to be a Weibull distribution with parameters $\lambda = 0.1$ and $\kappa = 2$. Simple accelerated median residual life model was assumed to include one binary covariate that was generated from a Bernoulli distribution with probability of success 0.5. Different censoring proportions were considered – 0%, 10%, 20% and 30% censoring. Sample sizes were *n* = 50, 100, 150 and 200 cases. For each numerical study we simulated *n* observations from the Weibull distribution with vector of parameters $\left(\frac{\lambda}{\exp(\beta_1 X_{1i})}, \kappa\right)$, $i = 1,..,n$ using the probability integral transformation technique (Casella and Berger, 2002) as

$$T_i = \frac{\exp(\beta_1 X_{1i})}{\lambda}(-\ln u_i)^{1/k}, \quad i = 1,..,n,$$

where u_i is from the uniform distribution between 0 and 1. The censoring times C_i's were generated from the uniform distribution between 0 and c, where c is a constant that controls for the censoring proportion. Then the observed data were determined as $Y_i = \min(T_i, C_i)$ and

$\delta_i = I(T_i \le C_i)$. We evaluate the empirical distribution of the regression parameter via sample mean and standard deviation based on 1000 simulations for each simulated dataset (*Table* 4-1). Sample average of the estimate of the location parameter κ of the baseline Weibull distribution across the 1000 simulations varied from 2.02 to 2.12 for all combinations of sample sizes and censoring proportions and sample average of the estimate of the scale parameter λ was approximately 0.10 across all scenarios. *Table 4-2* summarizes the estimated probabilities of Type I error for testing the null hypothesis $H_0 : \beta_1 = 0$. For the purpose of investigating the probabilities of Type I error, samples were generated under the null hypothesis $H_0 : \beta_1 = 0$. Distribution of power is presented in the *Table 4-3* for the case of sample size of 200 and different alternative values for β_1. For this part of our numerical studies samples were generated from the distributions with regression parameter β = 0.10, 0.15, 0.20, 0.25, and 0.30.

Table 4-1 Accelerated MERL model (parameter estimation bias and standard errors)

	Average censoring proportion							
	0%		10%		20%		30%	
n	$\Delta\beta_1$	SE	$\Delta\beta_1$	SE	$\Delta\beta_1$	SE	$\Delta\beta_1$	SE
50	0.0059	0.142	0.0003	0.152	0.0006	0.161	0.0088	0.178
100	0.0001	0.103	- 0.0036	0.109	-0.0009	0.112	-0.0030	0.125
150	-0.0003	0.082	0.0000	0.085	-0.0042	0.090	0.0002	0.101
200	0.0018	0.070	0.0023	0.075	-0.0004	0.080	-0.0009	0.082

Table 4-2 Accelerated MERL model (probability of type I error)

n	Average censoring proportion 0%	10%	20%	30%
50	0.059	0.064	0.060	0.066
100	0.060	0.063	0.055	0.066
150	0.055	0.058	0.052	0.059
200	0.053	0.050	0.053	0.050

Table 4-3 Accelerated MERL model (power, n = 200)

β	Average censoring proportion 0%	10%	20%	30%
0.10	0.317	0.286	0.266	0.197
0.15	0.578	0.509	0.497	0.422
0.20	0.821	0.745	0.711	0.668
0.25	0.945	0.936	0.889	0.860
0.30	0.990	0.978	0.973	0.947

The estimates of the regression coefficients are approximately unbiased and the corresponding standard errors show a systematic trend of increase with higher censoring proportion and decrease with larger sample size. Type I error probabilities are close to the prespecified level of

5%, and vary from 0.050 to 0.066 across all simulation scenarios with proximity to the 0.05 level when sample size increases. As it is expected, power decreases with higher censoring proportion and increase when the true value of the regression parameter moves away from the null value.

4.3 ACCELERATED MERL MODEL UNDER THE AFT ASSUMPTION

Another approach to avoid the difficulties related to the non-uniqueness of the survival distribution, when the median residual life function is known, is to restrict modeling to a family of survival distributions that is related in a certain manner specified a priori. We demonstrate that the accelerated failure time model provides such relationship between the survival functions that leads to the accelerated median residual life model.

Let's assume that the AFT model with acceleration factor $\rho = \exp(\boldsymbol{\gamma}'\mathbf{X})$ is satisfied

$$S(t) = S_0(\rho t),$$

then the following relationship between the inverse survival functions is also true

$$S^{-1}(y) = \frac{1}{\rho} S_0^{-1}(y).$$

Using these two equations the following set of relationships can be derived:

$$\begin{aligned}
\theta(t) &= S^{-1}(\tfrac{1}{2}S(t)) - t \\
&= \tfrac{1}{\rho} S_0^{-1}(\tfrac{1}{2}S(t)) - t \\
&= \tfrac{1}{\rho} S_0^{-1}(\tfrac{1}{2}S_0(\rho t)) - t \\
&= \tfrac{1}{\rho} \left\{ S_0^{-1}(\tfrac{1}{2}S_0(\rho t)) - \rho t \right\} \\
&= \frac{1}{\rho}\theta_0(\rho t)
\end{aligned}$$

Thus, as the functional form of the model we proposed is $\theta(t) = \eta\theta_0(t/\eta)$, where $\eta = \exp(\boldsymbol{\beta}'\mathbf{X})$, if the accelerated failure time assumption is assumed to be true, the accelerated median residual life model is also satisfied with the parameters of acceleration ρ and η that are reciprocal of each other $\rho = 1/\eta$, or in terms of regression coefficients $\boldsymbol{\beta} = -\boldsymbol{\gamma}$. This implies that to obtain the estimates for the regression parameters for the accelerated MERL model, it is sufficient to get the estimates of the coefficients for the AFT model and multiply them by (-1).

Usually another form of the AFT model is used, which linearly relates the logarithm of time variables to covariates of interest and which has a form $\ln(T) = \mu + \boldsymbol{\alpha}'\mathbf{X} + \sigma W$. This is an equivalent form of the AFT model with appropriately defined parameters. In this case as $\boldsymbol{\alpha} = -\boldsymbol{\gamma}$ and $\boldsymbol{\beta} = -\boldsymbol{\gamma}$, the estimate of the accelerated MERL model $\hat{\boldsymbol{\beta}}$ equals $\hat{\boldsymbol{\alpha}}$.

Therefore if an investigator is willing to assume that the accelerated failure time model is an assumption supported by the data and wants to make inferences on relationship between the covariates and the MERL function, any existing method can be applied to estimate the regression coefficients of the AFT model and the regression coefficients of the accelerated MERL model are automatically obtainable.

If no assumptions are made for the parametric form of the baseline distribution, semiparametric methods can be used to obtain the parameter estimates of the model (Miller, 1976; Buckley and James, 1979; Koul, Susarla and Van Ryzin, 1981; Chatterjee and Mcleish, 1986; Heller and Simonoff, 1990; Ritov, 1990; Tsiatis, 1990; Lai and Ying, 1991a, 1991b; Jin, Lin and Ying, 2006). Large sample properties of the parameter estimate for the accelerated MERL model $\hat{\boldsymbol{\beta}}$ would depend upon the properties of the parameter estimate from the AFT model $\hat{\boldsymbol{\alpha}}$.

The semiparametric method of Buckley and James (1979) is an extension of the least square method to fit the regression models for survival data. Since censored observations preclude the use of the regular least square method for parameter estimation for survival data, Buckley and James used an iterative procedure to estimate the regression parameters. This method has been shown to be superior to other extensions of the least square approaches to censored data (Lai and Ying, 1991a). The major difficulty in applying this or any other semiparametric method in practice is lack of software to perform the analysis. Recently, Stare, Harrell and Heinzl (2001) introduced an S-Plus program that allows for estimating the regression parameters using the Buckley and James method.

4.4 EXAMPLE

To illustrate two estimation techniques for the accelerated median residual life model – under the parametric assumption and AFT assumption – we simulated one sample dataset from a Weibull distribution and applied the proposed methods to this dataset. We assumed a simple regression with one binary covariate, which randomly divides the data between group 0 and group 1 in our notations. We generated a dataset of sample size 1000 with approximately 10% censoring proportion. Parameters of the Weibull distribution were assumed to be $\lambda = 0.1$ and $\kappa = 2$, and the true regression parameter β in the accelerated MERL model and therefore the regression parameter α in the AFT model were assumed to be equal to 0.4. We generated the data using the probability integral transformation technique described earlier in the text.

The maximum likelihood estimation technique was used to estimate the regression coefficients and their corresponding standard errors under the parametric assumption for the

baseline group. The estimates were $\hat{\lambda} = 0.099$, $\hat{\kappa} = 1.968$ and $\hat{\beta} = 0.418$ with relatively small bias for all parameters. The ML estimate of the standard error for parameter β was estimated to be equal 0.034, which gives a highly significant value of the Wald test statistic of 12.247. The comparison of true MERL functions, calculated using the corresponding formula for the Weibull distribution, nonparametric estimates and ML estimates of the median residual life functions in two groups is presented in *Figure 4-3*.

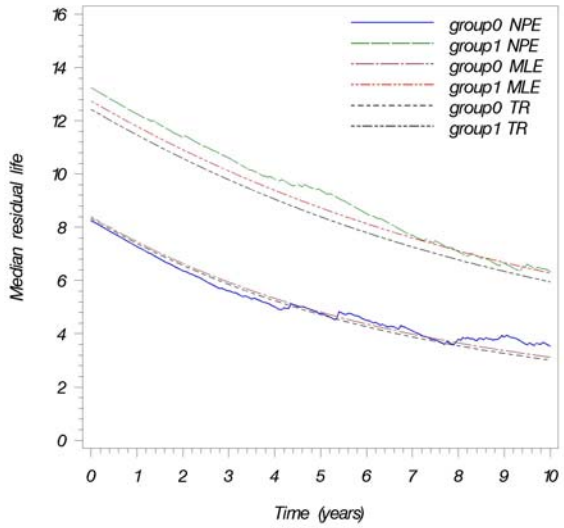

Figure 4-3 Accelerated MERL model (ML vs. nonparametric estimates)

As it is seen from the graph, all lines are very close to each other. The closeness of the true MERL functions and their parametric estimates was also evident from the estimated regression coefficients.

For semiparametric analysis of the same dataset, assuming that the accelerated failure time model is satisfied, we used Buckley and James method (BJ) to estimate the regression parameter and its standard error. The corresponding estimates were $\hat{\beta} = 0.415$ and $\widehat{SE} = 0.044$, which also produced a highly significant value of the Wald test statistic of 9.333. The comparison of nonparametric estimates and BJ estimates of the median residual life functions in two groups is presented in *Figure 4-4*.

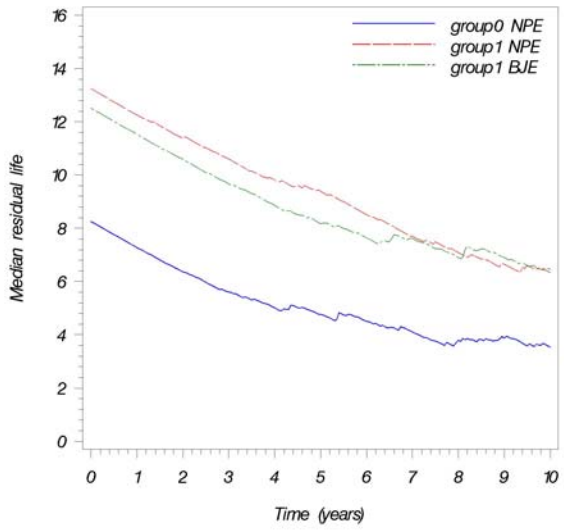

Figure 4-4 Accelerated MERL model (BJ vs. nonparametric estimates)

As it was expected, the closeness of the nonparametric curve and BJ estimate of the MERL function for group 1 is not as evident as in the parametric regression, though the Buckley and James method still provides a reasonable estimate.

In Figure 4-5 we combined all estimates described above.

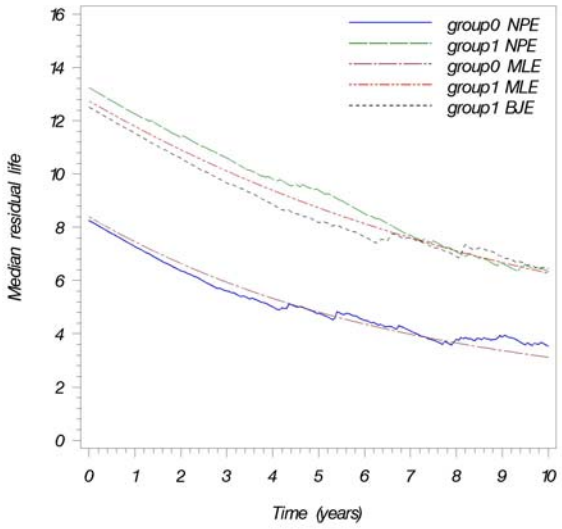

Figure 4-5 Accelerated MERL model (all curves combined)

4.5 SOME RELATIONSHIPS FOR THE MERL FUNCTIONS

4.5.1 Relationships under the accelerated MERL model

Suppose η is an acceleration factor for the accelerated median residual life model, which has the form $\theta(t) = \eta\theta_0(t/\eta)$. If we differentiate both sides of the equation with respect to *t*, we have $\theta'(t) = \theta_0'(t/\eta)$. As the derivative of the function at a point can be interpreted as the slope of the tangent line to the graph of the function at that point, this equation indicates that in the simple regression case the median residual life functions are "parallel" with a shift in the time axis.

Using the association between the derivatives of the median residual life functions, another interesting relationship between the derivatives of survival functions can be derived. The definition of the MERL function $\theta(t)$ gives $S(t+\theta(t)) = \frac{1}{2}S(t)$, and therefore by taking the derivative of both side of the equation, we get

$$S'(t+\theta(t))(1+\theta'(t)) = \frac{1}{2}S'(t), \quad \text{which implies} \quad \theta'(t) = \frac{S'(t)}{2S'(t+\theta(t))} - 1.$$

As $\theta(t) = \eta\theta_0(t/\eta)$ and $\theta'(t) = \theta_0'(t/\eta)$,

$$\frac{S'(t)}{S'(t+\theta(t))} = \frac{S_0'(t/\eta)}{S_0'(t/\eta+\theta_0(t/\eta))}$$

$$\text{which implies} \quad \frac{S'(t)}{S'(t+\eta\theta_0(t/\eta))} = \frac{S_0'(t/\eta)}{S_0'(t/\eta+\theta_0(t/\eta))}.$$

Now if we define $t_1 = t/\eta$ and $t_2 = t/\eta + \theta_0(t/\eta)$, then

$$\frac{S'(\eta t_1)}{S'(\eta t_2)} = \frac{S_0'(t_1)}{S_0'(t_2)} \quad \text{or} \quad \frac{S'(\eta t_1)}{S_0'(t_1)} = \frac{S'(\eta t_2)}{S_0'(t_2)} \quad \forall t_2 = t_1 + \theta_0(t_1).$$

Similar association between the survival functions can also be derived under the accelerated median residual life model. From the definition of the MERL function $\frac{S(t)}{S(t+\theta(t))}=2 \quad \forall t \geq 0$ and therefore this formula can also be applied to the baseline group for time point t/η as follows $\frac{S_0(t/\eta)}{S_0(t/\eta+\theta_0(t/\eta))}=2$. Equality of the right sides of the equations implies the equality of the left sides of these equations:

$$\frac{S(t)}{S(t+\theta(t))}=\frac{S_0(t/\eta)}{S_0(t/\eta+\theta_0(t/\eta))}$$

$$\text{which implies} \quad \frac{S(t)}{S(t+\eta\theta_0(t/\eta))}=\frac{S_0(t/\eta)}{S_0(t/\eta+\theta_0(t/\eta))}.$$

Now if we define $t_1=t/\eta$ and $t_2=t/\eta+\theta_0(t/\eta)$ as before, then

$$\frac{S(\eta t_1)}{S(\eta t_2)}=\frac{S_0(t_1)}{S_0(t_2)} \quad \text{or} \quad \frac{S(\eta t_1)}{S_0(t_1)}=\frac{S(\eta t_2)}{S_0(t_2)} \quad \forall t_2=t_1+\theta_0(t_1).$$

Therefore for any fixed time point $t_0 \geq 0$ and a corresponding set of points defined recursively as $A_{t_0}=\{t_i:\ t_{i+1}=t_i+\theta_0(t_i) \quad i=0,1,..\}$ the following is true:

$$\frac{S(\eta t_j)}{S_0(t_j)}=\frac{S(\eta t_k)}{S_0(t_k)} \quad \forall t_j,t_k \in A_{t_0}$$

If the initial point t_0 is chosen to be 0, then by definition of the survival function $S(\eta t_0)=S_0(t_0)=1$ and therefore the corresponding set A_0 possesses the accelerated failure time property of $S(\eta t_j)=S_0(t_j) \quad \forall t_j \in A_0$. Therefore the accelerated MERL model has a one-to-one correspondence with the AFT model at a specific set of points.

4.5.2 Relationship under the Cox proportional hazards model

The Cox proportional hazards model also induces a certain relationship between the percentile residual life functions. However the Cox model and the proportional MERL model are not as conjugate as the accelerated failure time and the accelerated median residual life models.

Let's assume that the Cox proportional hazards model is satisfied. Then we have

$$S(t) = S_0(t)^{\rho} \quad \text{or} \quad h(t) = \rho h_0(t)$$

and the following relationship linking the inverse survival functions is also true:

$$S^{-1}(y) = S_0^{-1}(y^{1/\rho}).$$

Using the definition of the MERL function and applying the above two equalities, we get the following relationship

$$\begin{aligned}
\theta(t) &= S^{-1}(\tfrac{1}{2}S(t)) - t \\
&= S_0^{-1}((\tfrac{1}{2}S_0(t)^{\rho})^{1/\rho}) - t \\
&= S_0^{-1}((\tfrac{1}{2})^{1/\rho}S_0(t)) - t \\
&= S_0^{-1}(pS_0(t)) - t \\
&= \theta_0^p(t).
\end{aligned}$$

Therefore $\theta(t) = \theta_0^p(t)$, where $p = (1/2)^{1/\rho}$. Here $\theta^p(t)$ defines a p^{th}-percentile residual life function, which, by the definition can be calculated as $\theta^p(t) = S^{-1}(pS(t)) - t$, since by definition $\theta^p(t)$ is such that $P(T - t > \theta^p(t) \mid T > t) = p$ or $S(t + \theta^p(t)) = pS(t)$.

4.6 DISCUSSION

In this chapter of the dissertation we have defined the accelerated median residual life model. This model is a functional analog to the accelerated failure time model. We proposed two methods of estimation of the regression coefficients. The first one is an example of the parametric regression model and assumes that the baseline distribution is known and it has a prespecified parametric form. For this situation the maximum likelihood estimation approach can be used to obtain the estimates of the regression coefficients and their standard errors. The second method assumes a specific relationship between the survival functions, i.e. the accelerated failure time assumption, which technically allows for both nonparametric and semiparametric estimation of the regression coefficients. We used the Buckley and James method as an example of the semiparametric estimation in this case.

The accelerated median residual life model presents another novel approach to model the relationship between the median residual life function and covariates of interest at multiple time points simultaneously. One of its main advantages is that most of the known parametric distributions, which are commonly used in the survival analysis, guarantee the uniqueness of the survival and MERL functions within that family of distributions, providing a great amount of flexibility for the model fit to the data. Also the relationship between the accelerated failure time model and accelerated median residual life model presents a simple way of drawing a conclusion about the median residual life function. Since we believe that the median residual life function can be of great value and importance in clinical research, this connection between two models will provide a useful way of describing the relationship between the MERL function and covariates if it is reasonable to assume that the accelerated failure time assumption is supported

by the data. Also the accelerated MERL model has a one-to-one correspondence with the AFT model at a specific set of points

On the other hand the accelerated MERL model is not as easy to interpret, as some other well known models or the proportional median residual life model. Though the relationship we described in section 4.5.1 may be helpful in providing some graphical explanation of this model.

The issues that arise due to a high censoring proportion also are relevant to this model as to the proportional median residual life model.

5.0 DISCUSSION AND FUTURE RESEARCH

Regression techniques are popular methodologies, especially in the field of survival analysis. It is of great importance to be able to describe the relationship between the covariates of interest, such as treatment, gender or age and some well-defined survival outcome, such as survival time or hazard function. The main idea of this dissertation was to develop two novel regression approaches that could model the relationship between the residual failure time distribution, represented by the median residual life function and a set of covariates. To our knowledge, the two proposed regression methods are the only frequentist models that attempt to model the median residual life function at multiple time points simultaneously and without any restrictions to a specific class of family distributions. The available methods regress the MERL function on important covariates at a specific time point (Ying, Jung and Wei, 1995; McKeague, Subramanian, and Sun, 2001; Yin and Cai, 2005; Jeong, Jung and Bandos, 2007), are focused on a specific class of parametric distributions (Rao, Damaraju, and Alhumoud, 1993) or model the MERL function induced by the accelerated failure time assumption using the Bayesian approach (Gelfand and Kottas, 2003).

The proportional median residual life model is a functional analog to the Cox proportional hazards model. It assumes the constant proportionality of MERL functions over the interval of interest. For this model we presented the semiparametric approach for parameter estimation, which required the minimization of an estimating function. We performed numerical

studies to evaluate performance of these estimates. The bootstrap resampling technique was used to estimate the corresponding standard errors that can be used to obtain confidence intervals for parameters of interest or perform hypothesis testing.

Several improvements and future directions can be considered regarding the proportional median residual life model.

- Proofs have to be completed regarding consistency of the estimator and its asymptotic normality.
- We believe that the asymptotic normality of the estimating function (3.3) can also be proven. Then minimum dispersion statistic (Basawa and Koul, 1988) could be derived for hypothesis testing and constructing confidence interval as proposed in Ying et al. (1995) and Jeong et al. (2007). We believe that this would substantially decrease the amount of time required for estimation of the standard errors, which was achieved with the help of the bootstrap resampling technique in this dissertation.
- Other methods for finding the function minima could be considered over the grid search that was used in the current work.
- Another area of improvement could come from modifying the estimation technique in such way that this model could be fitted to the data with a high censoring proportion.
- As the results of numerical investigations could depend on how the data were generated, it would be useful to find other distributions than exponential that possess the property of one-to-one correspondence between the MERL function and the survival function under the proportionality of the MERL functions assumption.
- The optimum choice of the interval of integration that is optimal in terms of the efficiency of the resulting regression estimator, the choice of the iteration scheme

described in section 3.1.2 and the number points required for the integral approximation are also among the future research topics.

- The problem of estimating the baseline median residual life function that arises with the presence of continuous covariates in the model should also be addressed in the future.

The accelerated median residual life model by its analytical form resembles the accelerated failure time model. For this model we presented two methods of estimation – parametric and semiparametric under the accelerated failure time assumption. Extensive numerical studies were carried out to evaluate the performance of the regression coefficient estimates under the parametric assumption. To illustrate how these methods work in practice one data realization was simulated from a Weibull distribution.

For this regression technique it would be desirable to come up with a semiparametric method of estimating the regression coefficients, which would not place any restrictions on the baseline MERL function, as in the parametric setting, or would not assume any specific relationship between survival functions, as in case of the AFT assumption.

For both models that were presented it would be advantageous to develop diagnostic methodology and techniques of model selection.

Considering the fact that the median residual life function is a special case of the quantile residual life function, similar regression models can be constructed to relate the quantile residual life function to the specified set of covariates, though appropriate changes have to be made.

BIBLIOGRAPHY

Alam, K., and Kulasekera, K.B. (1993), "Estimation of the quantile function of residual life time distribution," *Journal of Statistical Planning and Inference*, **37**, 327-337

Aly, E.A.A. (1992), "On some confidence bands for percentile residual life functions," *Nonparametric statistics*, **2**, 59-70

Arnold, B. C., and Brockett, P.L. (1983), "When does the βth percentile residual life function determine the distribution," *Operations Research*, **31**, 391-396

Barabas, B., Csörgö , M., Horvath, L., and Yandell, B.S. (1986), "Bootstrapped confidence bands for percentile lifetime," *Annals of the Institute of Statistical Mathematics*, **38**, 429-438

Basawa, I.V., and Koul, H.L. (1988), "Large-sample statistics based on quadratic dispersion," *International Statistical Review*, **56**, 199-219

Buckley, J., and James, I. (1979), "Linear regression with censored data," *Biometrika*, **66**, 429-436

Casella, G., and Berger, R.L. (2002), *Statistical Inference*, Pacific Grove: Duxbury

Chatterjee, S. and McLeish, D.L. (1986), "Fitting linear regression models to censored data by least squares and maximum likelihood methods," *Communication in statistics – Theory and Methods*, **15**, 3227-3243

Chung, C.F. (1989), "Confidence bands for percentile residual lifetime under random censorship model," *Journal of Multivariate Analysis*, **29**, 94-126

Cox, D.R. (1972), "Regression models and life-tables," *Journal of the Royal Statistical Society, Series B*, **34**, 187-220

Cox, D.R. (1975), "Partial likelihood," *Biometrika*, **62**, 269-276

Csörgö, M., and Csörgö, S. (1987), "Estimation of percentile residual life," *Operations Research*, **35**, 598-606

Csörgö , S., and Viharos, L. (1992), "Confidence bands for percentile residual lifetimes," *Journal of Statistical Planning and Inference*, **30**, 327-337

Efron, B. (1967), "The two sample problem with censored data," *In Proceedings of the Fifth Berkley Symposium on Mathematical Statistics and Probability*, New York: Prentice-Hall, **4**, 831-853

Efron, B. (1981), "Censored data and the bootstrap," *Journal of the American Statistical Association*, **76**, 312-319

Feng, Z., and Kulasekera, K.B. (1991), "Nonparametric estimation of the percentile residual life function," *Communication in Statistics: Theory and Methods*, **20**, 87-105

Fisher, B., Jeong, J., Anderson, S. et al. (2002), "Twenty-five-year follow-up of a randomized trial comparing radical mastectomy, total mastectomy, and total mastectomy followed by irradiation," *The New England Journal of Medicine*, **347**, 567-575

Gelfand, A.E., and Kottas, A. (2003), "Bayesian semiparametric regression for median residual life," *The Scandinavian Journal of Statistics*, **30**, 651-665

Ghosh J.K., and Mustafi C.K. (1986), "A note on the residual median process," *The Canadian Journal of Statistics*, **14**, 251-255

Gill, R.D. (1980), "Censoring and stochastic integrals," *Mathematical Centre Tracts*, Amsterdam: Mathematisch Centrum, 124

Gupta, R. C., and Langford, E.S. (1984), "On the determination of a distribution by its median residual life function: a functional equation," *Journal of Applied Probability*, **21**, 120-128

Haines, A.L., and Singpurwalla, N. D. (1974), "Some contributions to the stochastic characterization of wear," in *Reliability and Biometry*, 47-80, F. Proschan and R.J. Serfling (eds.). SIAM, Philadelphia

Heller, G. and Simonoff, J. S. (1990), "A comparison of estimators for regression with a censored response variable," *Biometrika*, **77**, 515-520

Jeong, J. (2006), "A new parametric family for modeling cumulative incidence functions: application to breast cancer data," *Journal of the Royal Statistical Society. Series A*, **169**, 289-303

Jeong, J., Jung, S.H. and Costantino, J. P. (2007), "Nonparametric inference on median residual lifetimes in breast cancer patients," *Biometrics*, published online

Jeong, J., Jung, S.H. and Bandos, H, (2007) "Regression on median residual life," *Journal of American Statistical Association: Theory and Methods*, the manuscript is in revision

Jin, Z., Lin, D.Y., and Ying, Z. (2006), "On least-squares regression with censored data," *Biometrika*, **93**, 147-161

Joe, H. (1985), "Characterizations of life distributions from percentile residual lifetimes," *Annals of the Institute of Statistical Mathematics*, **37**, 165-172

Joe, H., and Proschan, F. (1984), "Percentile residual life functions," *Operations Research*, **32**, 668-678

Johnson, N.L., and Kotz S. (1970) *Continuous Univariate Distributions*, I. John Wiley & Sons, New York

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

Klein, J.P. (1991), "Small-sample moments of some estimators of the variance of the Kaplan-Meier and Nelson-Aalen estimators," *Scandinavian Journal of Statistics*, **18**, 333-340

Klein, J.P. and Moeschberger, M.L. (2003), *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer

Koul, H., Susarla, V. and Van Ryzin, J. (1981), "Regression analysis with randomly right-censored data," *The Annals of Statistics*, **9**, 1276-1288

Lai, T.L. and Ying, Z. (1991a), "Large sample theory of a modified Buckley-James estimator for regression analysis with censored data," *The Annals of Statistics*, **19**, 1370-1402

Lai, T.L. and Ying, Z. (1991b), "Rank regression methods for left-truncated and right-censored data," *The Annals of Statistics*, **19**, 531-556

Lillo, R.E. (2005), "On the median residual lifetime and its aging properties:a characterization theorem and its applications," *Naval Research Logistics*, **52**, 370-380

McKeague, I.W., Subramanian, S. and Sun, Y. (2001), "Median regression and the missing information principle," *Nonparametric Statistics*, **13**, 709-727

Miller, R. (1976), "Least squares regression with censored data," *Biometrika*, **63**, 449-464

Rao, B.R., Damaraju, C. V., and Alhumoud, J. M. (1993), "Covariate effect on the life expectancy and percentile residual life functions under the proportional hazards and the accelerated life models," *Communication in Statistics: Theory and Methods*, **22**, 257-281

Ritov, Y. (1990), "Estimation in a linear regression model with censored data," *The Annals of Statistics*, **18**, 303-328

Schmittlein, D.C., and Morrison, D.G. (1981), "The median residual lifetime: a characterization theorem and application," *Operations Research*, **29**, 392-399

Song, J., and Cho, G. (1995), "A note on percentile residual life," *Sankhya*, **57**, 333-335

Stare, J., Harrell, Jr F.E., and Heinzl, H. (2001), "BJ: an S-Plus program to fit linear regression models to censored data using the Buckley-James method," *Computer Methods and Programs in Biomedicine*, **64**, 45-52

Tsiatis, A.A. (1990), "Estimating regression parameters using linear rank test for censored data," *The Annals of Statistics*, **18**, 354-372

Van der Vaart, A. W. (1998), *Asymptotic statistic (Cambridge series in statistical and probabilistic mathematics)*, Cambridge University Press

Yin, G. and Cai, J. (2005), "Quantile regression models with multivariate failure time data," *Biometrics*, **61**, 151-161

Ying, Z., Jung, S.H., and Wei, L.J. (1995), "Survival analysis with median regression model," *Journal of the American Statistical Association: Theory and Methods*, **90**, 178-184