The International Journal of Biostatistics

Volume 3, Issue 1

2007

Article 5

Regression Analysis of Mean Lifetime: Exploring Nonlinear Relationship with Heteroscedasticity

Zhiping Sun, *Merck & Co.*

Wenxin Jiang, *Northwestern University*

Recommended Citation:

Sun, Zhiping and Jiang, Wenxin (2007) "Regression Analysis of Mean Lifetime: Exploring Nonlinear Relationship with Heteroscedasticity," *The International Journal of Biostatistics*: Vol. 3: Iss. 1, Article 5.

DOI: 10.2202/1557-4679.1035

Regression Analysis of Mean Lifetime: Exploring Nonlinear Relationship with Heteroscedasticity

Zhiping Sun and Wenxin Jiang

Abstract

As a generalization of the accelerated failure time models, we consider parametric models of lifetime Y , where the conditional mean $E(Y|X; \beta)$ can depend nonlinearly on the covariates X and some parameters β . The error distribution can be heteroscedastic and dependent on X . With observed data subject to right censoring, we propose regression analysis for β based on Kaplan-Meier estimates of the means over several regions of X . Consistency and asymptotic distributional properties of the estimators are established under general conditions. A resulting estimator of β is shown to be the sum of two possibly dependent asymptotic normal quantities, based on which conservative confidence intervals and tests are derived. Simulation studies are conducted to investigate the performance of the proposed estimator and to compare it with Buckley-Jame's method. To illustrate the methodology, we study an example with kidney transplant data, where a nonlinear relationship called "mixtures-of-experts", proposed in the neural networks literature, is used to model the relationship between the survival time and the age of the patients.

KEYWORDS: accelerated failure time model, asymptotic normality, heteroscedasticity, Kaplan-Meier estimate, mixtures of experts, nonlinear regression

Author Notes: The authors are grateful to Tom Severini and Martin Tanner for useful comments.

1. INTRODUCTION

The accelerated failure time model (AFT) assumes $Y = X'\beta + e$, where Y is the lifetime or some transformation of it (e.g. logarithm of lifetime), $X' = (X_1, \dots, X_q)$ is the vector of covariates, β is the corresponding regression parameter, and e is the error having a density $f(e)$. The AFT model postulates a direct relationship between lifetime and covariates. This natural type of regression relationship led Sir David Cox (Kalbfleisch and Prentice 2002, p. 219) to remark that “accelerated life models are in many ways more appealing than the proportional hazards model because of their quite direct physical interpretation.” When the error distribution is unspecified, the AFT model can be considered as a semiparametric alternative to the relative risk or Cox models. Despite the related numerical challenges, the AFT model as a semiparametric family has received considerable attention in the statistical literature over the past decade or so.

This paper studies a generalization of the standard AFT. The AFT assumes a linear mean function and a common error distribution (e independent of X), while neither assumption is valid generally in practice. The purpose of this paper is to study an inferential method and apply it to a more general parametric model, where the mean lifetime can be modeled nonlinearly, and the inferential results remain valid even when e depends on X .

The model only specifies that the mean follows a parametric model (M): $E(Y|X) = g(X; \beta)$, where g is a pre-specified function, possibly nonlinear, and β is an unknown vector of parameters. No assumption is made on how the error $e = Y - E(Y|X)$ should depend on X . This allows the model to be used in situations with heteroscedasticity [nonconstant $\text{var}(Y|X)$].

In the case without censoring, estimation of β in the mean model (M) can be carried out with nonlinear least squares, based on minimizing the squared differences of observed Y 's and their parametric means. With right censoring, however, lifetime Y is not always observed, and the least squares method cannot be applied directly. Instead, we use a variant of least squares described below, where estimates of some ‘regional means’ are used in place of the possibly unobserved Y 's.

The main idea of the proposed estimation method involves dividing the domain of covariates X into a number of regions. In each region, we obtain two estimates of the ‘regional mean’: the mean estimated based on a Kaplan-Meier procedure, and the mean estimated from the parametric regression model (M). Then we minimize a possibly weighted sum of the squared

differences.

Our proposed estimation method is simple, intuitive and has explicit form in the linear case, while allowing for more general cases with nonlinearity and heteroscedasticity. It also allows for more relaxed assumptions on censoring than some other existing methods.

In Section 1.1, a brief literature review of some existing works is given. In Section 1.2, the rationale of modeling the mean life time for censoring data is discussed. The proposed mean lifetime regression model and the main idea of the estimation method are presented in Section 2. Section 3 presents theoretical results on asymptotic properties of the estimators. We will show that the resulting estimator is consistent and can be expressed as a sum of two possibly dependent asymptotic normal quantities. This can be used to construct conservative confidence intervals and tests. Section 4 studies numerically the performance of the proposed method, regarding how it responds to the number of regions used in the partition, the amount of censoring, etc. Section 5 illustrates the method using a real data set with kidney transplant patients, where a nonlinear model called ‘mixtures of experts’, developed in the neural networks literature (Jacobs et al. 1991), is used to model the mean lifetime. Brief conclusions are presented in Section 6. Conditions and proofs of the theoretical results are included as an Appendix.

Before proceeding we will introduce the framework and some notation. Assume that random vectors (X_i, Y_i, C_i) , $i = 1, \dots, n$, are independent and identically distributed (iid), where X denotes a vector of covariates, and Y denotes the survival time or some transformation of it with cumulative distribution function (CDF) F , and C denotes the censoring time with distribution function G . With right-censoring we only observe $Z_i = Y_i \wedge C_i$ and $\delta_i = I[Y_i \leq C_i]$.

1.1 Related Works

AFT and related models have been studied extensively in the literature. However, most existing methods assume common error and/or linear model for the mean log lifetime, or require relatively strong assumptions on the censoring mechanism; we will discuss these in detail below. Our method, on the other hand, handles nonlinearity and heteroscedasticity, and requires milder assumptions on the censoring mechanism.

Many existing estimation methods assume a common error distribution (e independent of X). Examples include rank-based methods (e.g., Kalbfleisch

and Prentice 2002, chap. 7), some least-squares based methods (e.g., Miller 1976, Buckley and James 1979), and some Bayesian methods (e.g., Kuo and Mallick 1997, Kottas and Gelfand 2001, Hanson and Johnson 2002). Actual errors may be dependent on X , e.g., through heteroscedasticity (different variances). Ignoring heteroscedasticity may not be a big problem when there is no censoring. For example, inferential results based on least squares remain consistent even though efficiency can be affected. This is not so when there is censoring. For example, Buckley and James (1979) use pooled errors to impute the unobserved Y 's, which are only known to exceed some censoring time C . The pooled errors, however, can have a variance being different from that of the actual errors, which are heteroscedastic. This can distort the imputed Y -values and lead to inconsistent inferential results. Our method, on the other hand, will be shown to be consistent under conditions that do not require a common error distribution.

Several other works also do not require a common error distribution. Koul, Susarla and Van Ryzin (1981)(KSV) use a synthetic data estimator and obtain an explicit and simple form in the linear case. Although neither KSV or Leurgans' (1987) synthetic data methods require homoscedasticity, they assume that the censoring distribution is the same for all values of X . The problem can be partially solved by stratification (see Fyngenson and Zhou 1994), assuming that C is independent of (Y, X) in each strata. A drawback of synthetic data methods is that the weights used in synthetic data can become large near the tail of the censoring distribution, possibly leading to inefficiency and unstable results.

Another modification of the KSV estimator, besides the stratification, is to generalize it to M-estimation (see Zhou 1999). Stute (1999) adapts this M-estimation procedure to non-linear models with right censoring. However, it is assumed that censoring time C and the lifetime Y are independent *unconditionally*. This is not very satisfactory, since in some likely situations, Y and C can both increase with X and induce an unconditional dependence. Such an assumption is relaxed in our paper, which only needs independence of Y and C within each stratification region based on X , which is easier to be approximately satisfied (see Remark 8 in Appendix A).

Akritis (1996) uses nonparametric kernel estimates for fitting polynomial regression models to a variant of the trimmed mean of Y given X . Akritis' method can accommodate heteroscedastic error, but is limited to models that are parameterized linearly. We, on the other hand, focus on more general situations where the mean function can depend on the parameters nonlinearly.

1.2 Modeling the mean life time for censoring data

There is some intrinsic difficulty in modeling the mean life time for censored data. Without some assumption it is an impossible task, since we would have no idea how much longer the censored observations will live (say, 1 year longer or 100 years longer), which will affect the mean estimate. (For a theoretical study on the bias of Kaplan-Meier mean estimates, see e.g., Stute 1994.) Although this has led to much research on alternative methods such as censored median regression (e.g., Ying, Jung and Wei 1995, Yang 1999), works on mean (or restricted mean) life times still remain a popular research direction (see, e.g., Akritas 1996, Gross and Lai 1996, Stute 1999). This is perhaps due to the fact that the mean incorporates information from all subjects, while the median only relies on the middle one or two observations and does not show the overall performance. Even if a treatment improved all lifetimes but the middle one, the median would not show the overall improvement while the mean would.

The contribution of the current paper is to show that under some assumptions (in Appendix), including a condition on the censoring distribution (Tail Condition), reliable inference on the mean model is still possible for nonlinear regression with heteroscedasticity. Even when the condition does not hold, the proposed method can still be interpreted as effectively modeling some kind of restricted mean (over the observed data range).

A related comment is that the current method, based on the mean estimation method by Gill (1983), can under-estimate the mean lifetime, especially when the last observation is censored. As a referee commented, the estimated regression effect of a covariate may still be relatively reliable if the underestimation occurs parallelly for all levels of the covariate. In practice, one can perform some kind of sensitivity analysis, possibly based on different ways of handling the censored last observations (e.g., Hosmer and Lemeshow 1999, Chapter 2), to see how the regression effects change. We will leave this as a subject of future research since it is outside of the main focus of the paper.

2. MODEL AND METHOD

The mean lifetime regression model we propose is

$$(M) : E(Y|X) = g(X; \beta)$$

where g is a known function and β is a q -dimensional coefficient vector.

The main idea of the proposed estimation method is to divide the domain of covariates X in m regions, $R_1, \dots, R_m$; in each region estimate the mean lifetime by the nonparametric method (Kaplan-Meier) and by the parametric method with the specified regression model; and then minimize the weighted sum of the squares of the differences. Thus, the estimator $\hat{\beta}$ is defined as:

$$\hat{\beta} = \arg \min_{b \in \Theta} \sum_{k=1}^m w_k \hat{\delta}_k^2(b),$$

where w_k 's are some positive weights, $\Theta \subset \mathcal{R}^q$ is a compact parameter space, and

$$\hat{\delta}_k(b) = \hat{E}_{KM}(Y|X \in R_k) - \frac{1}{n_k} \sum_{i: X_i \in R_k} g(X_i; b).$$

Here $\hat{E}_{KM}(Y|X \in R_k)$ denotes a Kaplan-Meier estimate of mean lifetime in region R_k . Although there are other possibilities (e.g., Stute 1995, Akritas 2000), we will use a commonly-used mean estimator by Gill (1983), where $\hat{E}_{KM}(Y|X \in R_k)$ is chosen to be $\hat{\mu}_k^{(n_k)} = \int_0^{Z_{(n_k)}} \{1 - \hat{F}_k(u)\} du$, and $\hat{F}_k$ is the Kaplan-Meier estimator of the CDF of Y , based on data with X -values falling in region k . Here n_k is the number of X 's in region R_k , $Z_{(n_k)}$ is the largest observed event time in region k , $k = 1, \dots, m$.

The weight w_k 's can be taken simply as 1. We use a general framework where w_k 's can vary, since varying w_k 's may lead to improved efficiency in parameter estimation (see Remark 2 later).

Our proposed estimator has an explicit form in the linear case, since the objective function would then be quadratic in b . On the other hand, our method allows more general nonlinear mean functions. Another advantage of this generalized model is that it does not require the errors to be identically distributed. Since this is a distribution-free model, it allows the distribution of Y given X to have different shapes for different X values and it allows heteroscedasticity. .

3. THEORY

3.1 Main results

In order to find the asymptotic distribution of the estimators, first we look at the case in which the distribution of X is known, possibly obtained from other

studies. In this scenario we can compute the expected lifetime in each region directly with the specified $g(\cdot)$ function in the regression model, and compare it to the Kaplan-Meier mean estimate. Define $\tilde{\beta} = \arg \min_{b \in \Theta} \sum_{k=1}^m w_k \tilde{\delta}_k^2(b)$, where $\tilde{\delta}_k = \hat{E}_{KM}(Y|X \in R_k) - E(g(X; b)|X \in R_k)$, and $\Theta \subset \mathcal{R}^q$ is a compact parameter space. To derive the asymptotic distribution of $\tilde{\beta}$, we will use the fact that under some mild regularity conditions, $\hat{E}_{KM}(Y|X \in R_k)$ has an asymptotic normal distribution with asymptotic variance $n_k^{-1}\sigma_k$ for some $\sigma_k > 0$, as $n_k \rightarrow \infty$. (See Remark 7 in Appendix A.)

We will use $\partial_b f(b)$ to denote the column vector of derivatives, and $\partial_b^2 f(b)$ to represent the second order derivative matrix, for a function $f(b)$. Then the following result holds:

Theorem 1 (With known X -distribution). Assume that the conditions 1-5 in Appendix A hold. Then there exists a parameter space Θ being a compact neighborhood of the true parameter, over which the minimizer $\tilde{\beta} = \arg \min_{b \in \Theta} \sum_{k=1}^m w_k \tilde{\delta}_k^2(b)$ is $\sqrt{n}$ -consistent for the true parameter β , and that $\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, V_1)$, as $n \rightarrow \infty$. Here $V_1 = A^{-1}B_1A^{-1}$,

$$A = \partial_b^2 \sum_{k=1}^m w_k [E(Y|X \in R_k) - E(g(X; b)|X \in R_k)]^2 \Big|_{b=\beta},$$

$$B_1 = 4 \sum_{k=1}^m w_k^2 (\sigma_k / \pi_k) \partial_b E(g(X; b)|X \in R_k) \partial_b' E(g(X; b)|X \in R_k) \Big|_{b=\beta},$$

and $n_k^{-1}\sigma_k$ is the asymptotic variance of $\hat{E}_{KM}(Y|X \in R_k)$.

When the distribution of X is unknown, we can use the estimator $\hat{\beta}$ defined as in Section 2, where the data in region R_k , $k = 1, \dots, m$, are used to estimate the Kaplan-Meier mean of lifetime as well as the mean of the $g(\cdot)$ function. Thus, in each region, the two mean estimators are not independent and the asymptotic distribution of the difference is unknown, since the joint distribution of the two mean estimators are unknown. However, the asymptotic distribution of $\hat{\beta}$ can still be shown to be a sum of two (possibly dependent) asymptotic normal quantities. This enables us to construct conservative confidence intervals and hypothesis tests to do the inference (Corollaries 1 and 2).

Theorem 2 (*With unknown X -distribution*). Assume that the conditions 1-5 in Appendix A hold. Then there exists a parameter space Θ being a compact neighborhood of the true parameter, over which the minimizer $\hat{\beta} = \arg \min_{b \in \Theta} \sum_{k=1}^m w_k \hat{\delta}_k^2(b)$ is $\sqrt{n}$ -consistent for the true parameter β , and that $\sqrt{n}(\hat{\beta} - \beta) = S_n + T_n$, where $S_n \xrightarrow{d} N(0, V_1)$ and $T_n \xrightarrow{d} N(0, V_2)$, as $n \rightarrow \infty$. Here $V_2 = A^{-1}B_2A^{-1}$,

$$B_2 = 4 \sum_{k=1}^m w_k^2 (\nu_k / \pi_k) \partial_b E(g(X; b) | X \in R_k) \partial'_b E(g(X; b) | X \in R_k) \Big|_{b=\beta},$$

$$\nu_k = \text{Var}(g(X; \beta) | X \in R_k).$$

Assume that the conditions 1-5 in Appendix A hold, so that the result of Theorem 2 holds. Then we have the following corollaries.

Corollary 1 (*Confidence Interval*). An asymptotically conservative $100(1 - \alpha)\%$ confidence interval for $C^T \beta$ (C is a q -dimensional constant vector) can be defined as

$$C^T \hat{\beta} \pm \left(\Phi^{-1} \left(1 - (1 - \eta) \frac{\alpha}{2} \right) \sqrt{C^T \frac{V_1}{n} C} + \Phi^{-1} \left(1 - \eta \frac{\alpha}{2} \right) \sqrt{C^T \frac{V_2}{n} C} \right),$$

i.e.

$$\liminf_{n \rightarrow \infty} Pr \left(|C^T (\hat{\beta} - \beta)| < \Phi^{-1} \left(1 - (1 - \eta) \frac{\alpha}{2} \right) \sqrt{C^T \frac{V_1}{n} C} + \Phi^{-1} \left(1 - \eta \frac{\alpha}{2} \right) \sqrt{C^T \frac{V_2}{n} C} \right) > 1 - \alpha,$$

where η can be any number between 0 and 1. (Of course, we can use numerical search method to find the optimal η such that the length of the confidence interval is the shortest.)

Corollary 2 (*Hypothesis Test*). Consider hypothesis testing at a significance level α : $H_0 : C^T \beta = b_0$ versus $H_1 : C^T \beta \neq b_0$, where C is a q -dimensional constant vector and b_0 is a constant. An asymptotically conservative rejection rule can be to reject H_0 whenever

$$|C^T \hat{\beta} - b_0| > \Phi^{-1} \left(1 - (1 - \eta) \frac{\alpha}{2} \right) \sqrt{C^T \frac{V_2}{n} C} + \Phi^{-1} \left(1 - \eta \frac{\alpha}{2} \right) \sqrt{C^T \frac{V_1}{n} C},$$

where η can be any number between 0 and 1. That is, under H_0 ,

$$\lim_{n \rightarrow \infty} \sup Pr \left\{ |C^T \hat{\beta} - b_0| \geq \Phi^{-1}(1 - (1 - \eta)\frac{\alpha}{2}) \sqrt{C^T \frac{V_2}{n} C} + \Phi^{-1}(1 - \eta\frac{\alpha}{2}) \sqrt{C^T \frac{V_1}{n} C} \right\} \leq \alpha.$$

(The optimal η found in Corollary 1 is also the one that gives the most powerful test.)

The proofs of these results are given in the Appendix.

We here assume that V_1 and V_2 are known. In practice they can be estimated consistently as discussed in Remark 1. Using these estimated versions will not affect the validity of inference on β .

3.2 Comments

Remark 1 (Variance estimates). The asymptotic variances can be estimated, by replacing the unknown true parameter by the estimated ones ($\tilde{\beta}$ or $\hat{\beta}$), replacing the expectations by the sample averages, replacing π_k by n_k/n , and replacing σ_k by a consistent estimate $\hat{\sigma}_k$ (e.g., the one discussed in Kumazawa 1987 taking $t = 0$), for all $k = 1, \dots, m$.

Remark 2 (Choosing weights). Any choice of the weights w_k 's can lead to consistent estimates. The simplest unweighted least squares with $w_k = 1$ for all k can be used. However, we note that it is possible to increase the efficiency, possibly in a 2-stage procedure, by choosing different w_k 's. It can be easily shown that the asymptotic variance of $\tilde{\beta}$ becomes minimized (in the matrix sense) by choosing $w_k = \pi_k/\sigma_k$, $k = 1, \dots, m$, if the mean model (M) is correctly specified. (These weights can be estimated according to the recipe of Remark 1.) Such a choice of weights will also approximately produce tightest asymptotic confidence intervals based on $\hat{\beta}$, if the second asymptotic variance V_2 is very small in the matrix sense. This happens if the partitions R_k 's are so chosen that ν_k 's (variance of $g(X; \beta)$ over $X \in R_k$) are very small compared to the σ_k 's (n_k times the asymptotic variance of the Kaplan Meier (KM) mean estimate), which happens when R_k 's are very small regions, or chosen such that the mean lifetime varies very little within each region.

Remark 3 (Choosing a partition). Very simple procedure for choice of R_k 's may work pretty well. When there are discrete components of X , the regions may be naturally chosen to take each possible value of that component.

For continuous component of X , one possibility is to choose the regions to approximately equalize the number of observations. This approach will be taken in the real data example and simulation experiments later.

The theoretical results above place some restrictions on the choice of the partition R_k 's. Condition 4 in Appendix A requires that m , the number of partitions, should be at least equal to the number of parameters q . This is useful for identifiability. On the other hand, Condition 3 in Appendix A implies that the number of partitions should be such that n_k , the number of observations in each region, should be increasing to infinity as n increases. This will be useful for the KM mean estimators and the resulting β estimates to be asymptotically consistent. No other restrictions are directly needed for our consistency results to hold. Simulation experiments in Section 4 seem to suggest a rule of thumb, that having 20 or more observations per region, with at least 5 to 10 of them uncensored, often generates satisfactory empirical results.

Remark 4 (About Theorem 2). We here use the Gill (1983) version of the KM estimates of the means, but in principle, any asymptotic normal region-specific mean estimates satisfying Condition 3 (in Appendix A) can be used. The resulting estimator $\hat{\beta}$, used when the distribution of X is unknown, is a sum of two possibly dependent asymptotic normals. Consequently only conservative confidence intervals and tests are derived as in the corollaries. This approach, however, often is not overly conservative, especially when the second asymptotic normal variance is much smaller. As described at the end of Remark 3: if the partitions R_k 's are so chosen that $g(X; \beta)$ varies very little over each region of X (perhaps by choosing R_k 's to be relatively small), the variation of the second asymptotic normal can be approximately ignored compare to the first one. Then one could approximately use the result of Theorem 1 for nonconservative inference, since now the estimators $\hat{\beta}$ and $\tilde{\beta}$ are about the same.

It is also noted that some other versions of the region-specific mean estimates can actually have an iid-sum asymptotic representation. If such mean estimates, instead of those based on Gill (1983), are used to produce $\hat{\beta}$, the resulting estimate is then one simple asymptotic normal quantity, instead of a sum of two. Such alternative mean estimates are discussed, e.g., in Stute (1995) and Akritas (2000).

Remark 5 (Model misspecification). We will mainly assume (M) to be the correct model for the mean. However, the inferential results still have mean-

ing even if model (M) is misspecified. The estimated parametric mean $g(x; \beta)$ then in some sense is estimating a best approximation to the true mean function $E(Y|X)$. With a fine partition of the regions and with weights w_k 's chosen to be π_k 's, approximately the best approximation in L_2 sense is estimated.

4. SIMULATIONS

The performance of the proposed method has been examined through simulation studies. In all the simulations, the covariate X has two independent components X_1 and X_2 : X_1 is uniformly distributed on $(0, 1)$; X_2 is binary with equal probability of being 0 or 1.

4.1 Linear Model

We first simulate data according to a linear model. In the simulation, the lifetime distribution $Y|X$ is normal with mean $1.5 + 0.5X_1 + 0.6X_2 + 0.4X_1X_2$ and variance $0.4X_1 + 0.4X_1X_2$. Since the variance depends on covariates X , this is a heteroscedastic example. The censoring distribution $C|X$ is normal with mean $\mu_c + 0.5X_1 + 0.6X_2 + 0.4X_1X_2$ and variance $0.6X_1 + 0.6X_1X_2$, where μ_c is used to adjust the percentage of censoring.

Table 1 shows the performance of the proposed method by listing the averages and standard deviations (in parentheses) of the estimated parameters, which are obtained from 100 simulated data sets of sample size $n = 200$ for each situation. A ten-region partition (corresponding to an average of 20 observations per region) and the unweighted approach are used in estimation with the proposed method. The table also shows the performance of the popular Buckley-James' (BJ) method using the free software 'bj' in R, which serves as a benchmark for evaluating the performance. We investigate the performance of both methods under different amounts of censoring.

1. When the censoring is not heavy (about 25%), the proposed estimator is less biased than BJ estimator with about the same standard deviation.
2. Under moderate censoring (50%), the proposed estimator is much less biased than BJ estimator with about the same standard deviation.

Table 1: Performance of the Proposed Method (Linear Model)

		β_0	β_1	β_2	β_{12}
True Value		1.5	0.5	0.6	0.4
Censoring	Method				
25% ($\mu_c = 2.0$)	Proposed	1.498 (0.068)	0.478 (0.153)	0.609 (0.139)	0.409 (0.322)
	BJ	1.525 (0.068)	0.456 (0.147)	0.622 (0.135)	0.318 (0.304)
50% ($\mu_c = 1.5$)	Proposed	1.503 (0.077)	0.453 (0.170)	0.595 (0.141)	0.404 (0.334)
	BJ	1.695 (0.094)	0.235 (0.160)	0.552 (0.149)	0.368 (0.324)
75% ($\mu_c = 1.0$)	Proposed	1.393 (0.151)	0.544 (0.298)	0.684 (0.252)	0.267 (0.479)
	BJ	1.967 (0.206)	-0.061 (0.309)	0.481 (0.237)	0.357 (0.447)

- Heavy censoring (75%) leads to higher standard deviations for both methods. The proposed method is still much less biased for most coefficients.

4.2 Nonlinear model

Now we consider a nonlinear model. The conditional distribution of lifetime Y given covariate vector $X = (X_1, X_2)$ is a gamma distribution with parameter λ and k , i.e.

$$f(y|x)_{Y|X} = \frac{\lambda(\lambda y)^{k-1} e^{-\lambda y}}{\Gamma(k)},$$

where

$$\begin{aligned}\lambda &= \exp(0.4x_1 + 0.1x_2 + 0.1x_1x_2 - 3), \\ k &= \exp(0.9x_1 + 0.4x_2 + 0.2x_1x_2 - 1.5).\end{aligned}$$

The conditional distribution of censoring time C given X is a Pareto distribution with parameter a and b , i.e.

$$f(c|x)_{C|X} = \frac{ab^a}{(c+b)^{a+1}}, c \geq 0,$$

where

$$a = 2,$$

$$b = \exp(0.3x_1 + 0.1x_2 + \gamma_c),$$

and γ_c controls the amount of censoring.

In this simulation set-up, heteroscedasticity is present since the conditional variance of lifetime is $Var(Y|X) = \exp(0.1X_1 + 0.2X_2 + 1)$, depending on the value of the covariates. Also, the true regression model proposes a nonlinear mean function

$$E(Y|X; \beta) = \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_0),$$

where $(\beta_1, \beta_2, \beta_{12}, \beta_0) = (0.5, 0.3, 0.1, 2.5)$. It can be shown that $E(\ln(Y)|X)$ is not a linear function of X or of the parameters β ; simply taking logarithm of the lifetime can not reduce it to a linear model. In fact, in simulation results not reported here, we applied the `bj` program in R, which is designed for linear models, to conduct a naive analysis assuming a linear relationship of $E(\ln(Y)|X)$. This led to a poor performance with the estimates being biased and having large variations, when there is moderate or heavy amount of censoring.

In simulations below, we resort to the proposed generalized lifetime regression model which can accommodate nonlinear mean functions as well as heteroscedasticity.

Table 2 shows the means and standard deviations (in parantheses) of the proposed estimates, based on 50 simulated data sets in each situation (with sample size $n = 200$). We investigate the performance of each method under different censorships. (Ten-region partition and unweighted approach were used in estimation with the proposed method.)

We notice that the estimated coefficients have means close to the true parameters in all three cases of censoring (25%, 50% and 75%). The standard deviations of the estimates, on the other hand, do increase slowly with increased amount of censoring. Overall, the results show that the proposed method has successfully handled both nonlinearity and heteroscedasticity.

Table 2: Performance of the Proposed Method (Nonlinear Model)

	β_1	β_2	β_{12}	β_0
True Value	0.5	0.3	0.1	2.5
Censoring				
25%	0.4923	0.2864	0.1046	2.5070
($\gamma_c = 4.5$)	(0.0440)	(0.0334)	(0.0542)	(0.0292)
50%	0.4840	0.2748	0.1158	2.5128
($\gamma_c = 3.6$)	(0.0546)	(0.0441)	(0.0646)	(0.0372)
75%	0.4799	0.2917	0.1169	2.5074
($\gamma_c = 2.8$)	(0.0788)	(0.0576)	(0.1050)	(0.0480)

4.3 Partitions and Weights

Now we investigate how the number of regions used to partition the covariate domain affects estimation with the proposed method.

The details of the simulations are described as follows. We simulate data according to the nonlinear gamma regression model as before, where $Y|X$ follows a gamma distribution with parameters λ and k ; $\lambda = \exp(0.4x_1 + 0.1x_2 + 0.1x_1x_2 - 3)$, $k = \exp(0.9x_1 + 0.4x_2 + 0.2x_1x_2 - 1.5)$. The censoring distribution $C|X$ is a Pareto distribution with parameters a and b , i.e. $f(c|x)_{C|X} = ab^a / (c+b)^{a+1}$, $c \geq 0$, where $a = 2$, $b = \exp(0.3x_1 + 0.1x_2 + 3.2)$. The covariate X has two independent components (X_1, X_2) , where X_1 is uniform on $(0, 1)$ and X_2 is binary with equal probability of being 0 or 1.

The sample size of each simulated data set is 200. Summary results of 50 runs are shown in Table 3, where “Mean” is the average of the estimates and “Std” is the standard deviation of the estimates.

The partition scheme is that, if the number of partitions is $m = 2j$, then for each value of X_2 (0 or 1), the domain of X_1 (i.e. $(0, 1)$) is partitioned into j equal-sized intervals. We have tried three cases $m = 4, 10, 30$, corresponding to about 50, 20, 7 observations per region, on average. The censoring distribution for this simulation setup generates about 65% censored observations. This implies that on average, the three cases have number of *uncensored* observations per region being 18, 7, 3, respectively. The case $m = 4$ uses the minimum number of regions needed to identify all 4 parameters. We see from Table 3 that for all three cases of m , the average values of the estimates are

Table 3: Parameter Estimation Using Different Number of Partitioned Regions (Unweighted and Weighted Method)

Parameter	True Value		Method	Number of Partitions		
				4	10	30
β_1	0.5	mean	un wt	0.4801	0.4867	0.4830
			wt	0.4803	0.4872	0.4794
		std	un wt	0.0512	0.0402	0.0575
			wt	0.0506	0.0400	0.0749
β_2	0.3	mean	un wt	0.2850	0.2876	0.2864
			wt	0.2852	0.2889	0.2622
		std	un wt	0.0493	0.0329	0.0416
			wt	0.0489	0.0318	0.0594
β_{12}	0.1	mean	un wt	0.1253	0.1161	0.1177
			wt	0.1250	0.1153	0.1353
		std	un wt	0.0696	0.0563	0.0643
			wt	0.0689	0.0552	0.0814
β_0	2.5	mean	un wt	2.5122	2.5094	2.5102
			wt	2.5120	2.5090	2.5216
		std	un wt	0.0333	0.0312	0.0401
			wt	0.0330	0.0306	0.0616

very close to the true values. On the other hand, the ten-region partition ($m = 10$) gives the smallest standard deviations of the estimates, followed next by the case of $m = 4$ (in most situations).

Table 3 also shows comparisons between weighted and unweighted least square estimates, where weights are based on the optimal weights (see Remark 2 in Section 3. For four-region and ten-region partition, the mean of the estimates from the weighted method is slightly closer to the true value than that from the unweighted method for each parameter, and the standard deviation of the estimates is slightly smaller from the weighted method. However, for thirty-region partition, such advantages of weighted method can not be found, the reason is that the weights are based on the asymptotic variance of the Kaplan-Meier mean lifetime in each region, in which there are too few observations and the asymptotic variance formula is not applicable.

Combining these findings, it seems that choosing a partition such that on

average there are at least 20 observations per region, with at least 5 to 10 of them uncensored, often gives satisfactory results, whether the unweighted or the weighted method is used. Using the weighted method (based on the optimal weights) can lead to some small improvement over the unweighted approach.

5. EXAMPLE

We illustrate the methodology using the Kidney Transplant Data from Chapter 1 of Klein and Moeschberger (1997).

The response variable Y will be taken as the (natural) logarithm of time to death; C is the logarithm of the censoring time. There are 863 patients and 140 of them are uncensored. We will study how Y is related to a covariate X , which is the age when transplant was performed on a patient.

Before performing data analysis, based on common-sense knowledge, we suspect that young patients and old patients may behave differently and have different aging effects on survival; young people may be still growing, while older people deteriorate more rapidly in body functions. To capture such differences in a smooth parametric way, we consider a ‘mixture of two experts’ model originated from neural networks literature (ME, Jacobs et al., 1991), originally of the form

$$(ME) : E(Y|X) = g(X; \beta) = \frac{e^{v(X-x_0)}}{1 + e^{v(X-x_0)}}(\alpha + \gamma X) + \frac{1}{1 + e^{v(X-x_0)}}(\alpha^* + \gamma^* X).$$

The term ‘experts’ actually refers to the standard linear models $(\alpha + \gamma X)$ and $(\alpha^* + \gamma^* X)$, which are proposed with changing mixing-weights dependent on X , the mixing-weights being, respectively, $e^{v(X-x_0)}/\{1 + e^{v(X-x_0)}\}$ and $1/\{1 + e^{v(X-x_0)}\}$. The location x_0 is the place where the mixing-weights of the two experts are equal. Suppose, for example, v is positive, then for $X > x_0$, the first expert $(\alpha + \gamma X)$ has a higher mixing-weight; for $X < x_0$, the second expert $(\alpha^* + \gamma^* X)$ has a higher mixing-weight.

This original form of ME model does not require that the two experts agree at the joint location x_0 . This sometimes can lead to abrupt changes near x_0 , which can lead to numerical instability in model-fitting.

We restrict the parametrization so that $\alpha + \gamma x_0 = \alpha^* + \gamma^* x_0$, by choosing $\alpha^* = \alpha - \delta x_0$, $\gamma^* = \gamma + \delta$. Then the model can be re-written as

$$(ME)' : E(Y|X) = g(X; \beta) = \alpha + \gamma X + \frac{1}{1 + e^{v(X-x_0)}}\delta(X - x_0),$$

Table 4: Mixture of Two Experts Parameter Estimation

Parameter	Estimate	95% CI	CI half width
α	7.884	(7.262, 8.506)	0.622
γ	0.0039	(-0.0218, 0.0296)	0.0256
v	-0.90	(-16.55, 14.75)	15.65
x_0	31.2	(23.6, 38.8)	7.6
δ	-0.0301	(-0.0528, -0.0074)	0.0227

where Y is log-lifetime, X is age, and the parameter β has components α , γ , v , x_0 and δ ; α and γ are the intercept and slope of the first expert, $1/\{1 + e^{v(X-x_0)}\}$ is the mixing-weight of the second expert, δ is the slope difference between the first and second expert, and x_0 is the age where mixing-weights of the first and second expert are equal (i.e., 0.5). In this restricted form (ME)', the values of the two experts are the same at x_0 , which forces a gradual and non-abrupt transition between the two linear models.

Ages of the patients range from 1 to 75. This domain is divided into 17 regions with approximately 50 observations in each region. The Kaplan-Meier (KM) mean for each region is estimated. The parameter estimates are obtained by minimizing a weighted sum of squared differences between the KM mean estimates and the parametric mean estimates based on (ME)'. The weights used in the weighted least squares are based on the optimal weights, i.e. the ratio of cell size to the variance of the KM mean of this cell (see Remark 2 in Section 3).

The estimated parameters and their conservative 95% confidence intervals are given in Table 4. The conservative confidence intervals are obtained based on Corollary 1, where a grid of η 's between 0 and 1, spaced 0.1 apart, is searched over to produce a tightest confidence interval for each parameter component.

We see that the confidence interval for δ , the slope difference of the two experts, does not contain a neighborhood of 0, which indicates a significant change of slope for effect of aging. The confidence interval for x_0 is centered around 31. The interpretation is that some change occurs around age x_0 : before that the survival time remains stable (and actually increases slightly) as a person's age increases; after that the survival time decreases much more rapidly with aging. This is also reflected in the fitted mean curve $g(x; \hat{\beta})$

Figure 1: Fitted Mean Log-lifetime and 95% Confidence Interval

Figure 2: Mixing-weights of the Experts and Their 95% Confidence Intervals

shown in Figure 1.

Figure 1 and Figure 2 provide the estimated curves and conservative 95% pointwise confidence bands, for the mean function and for the mixing-weights of the two experts, respectively. The conservative confidence intervals are obtained from applying Corollary 1, using linear approximations similar to the delta method (see, e.g., Lehmann 1991, Theorems 5.1.5, 5.1.9).

6. CONCLUSIONS

This paper provides a method of inference for a class of parametric mean lifetime models generalizing the standard accelerated failure time model. Although there exists extensive literature in this area, our approach stands out for its capability to model nonlinear relations with heteroscedasticity, while requiring milder assumptions on the censoring mechanism. The resulting estimator is shown theoretically to be consistent, and can be expressed as a sum of two asymptotic normal quantities which are possibly dependent. Conservative confidence intervals and tests are constructed.

The method of inference is based on mean estimates from partitioned

regions. One possible future project is to develop an automated procedure to choose those regions, possibly for increased efficiency. Some method similar to CART (Breiman et al. 1984) or survival trees (Breiman 2002) may be adapted for this purpose. Alternatively, one may consider inference based on overlapping regions, which would extend the approach of Akritas (1996) by using kernel estimates to fit *nonlinear* models.

We have considered minimization of a sum of squared differences between two estimates of regional means. Alternatively, one may minimize a sum of absolute values of these differences. This may relate to methods based on medians (e.g., Ying, Jung and Wei 1995, Yang 1999), which can bring desirable robustness properties, at the cost of numerical challenges of minimizing nondifferentiable functions.

APPENDIX

A. Regularity conditions

Define R_1^m as a mutually exclusive and exhaustive fixed partition of the domain of X . That is, $\bigcup_{k=1}^m R_k = \text{Dom}(X)$ and $R_j \cap R_k = \emptyset, \forall j \neq k$.

For $k = 1, \dots, m$, let $\mathcal{A}_k = \{i : X_i \in R_k, i = 1, 2, \dots\}$, which is a random object describing which elements of a sample path $\{X_1(\omega), X_2(\omega), \dots\}$ fall in region R_k . Let $\mathcal{A} = \mathcal{A}_1^m$ denote the collection of all these $\mathcal{A}_k$'s, which describes the random partition of all indexes of the sample path into the m regions. Let n_k be the number of X 's (from the n observed X 's) falling in region R_k . In terms of the cardinality of sets, $n_k = \text{card}\{i : X_i \in R_k, i = 1, 2, \dots, n\} = \text{card}(\mathcal{A}_k \cap \{1, 2, \dots, n\})$, for each $k = 1, \dots, m$. Some of the regularity conditions stated below will involve conditioning on a random partition $\mathcal{A}$. Our main results (stated in Section 3.1) will also be proved in Appendix B by first conditioning on a random partition $\mathcal{A}$.

The following regularity conditions will be used in proving the theorems.

Condition 1 (Domain of X). The domain of X , $\text{Dom}(X)$, is a compact set.

Condition 2 (Smoothness of parametrization). $\partial_b^2 g(x; b)$ is continuous in x and b for all $(x, b) \in \text{Dom}(X) \times \mathbb{R}^q$.

Condition 3 (Mean estimates). In each region, the Kaplan Meier estimator of the mean lifetime is $\sqrt{n_k}$ -consistent and has an asymptotic normal (AN) distribution as $n_k \rightarrow \infty$, conditional on a random partition $\mathcal{A}$.

Condition 4 (Local identifiability). The $m \times q$ matrix, whose (k, j) th element is $\partial_{b_j} E(g(X; \beta) | X \in R_k)$, $k = 1, \dots, m$, $j = 1, \dots, q$, has rank q . (This implicitly requires $m \geq q$.)

Condition 5 (Nonnull partitions). The regions R_1^m form a partition of $Dom(X)$, such that $\pi_k \equiv Pr(X \in R_k) > 0$, $\forall k = 1, \dots, m$.

Remark 6 (Conditions on domain of X and the nonnull partition). Although we have assumed in Conditions 1 and 5 that the domain of X is compact, and the partition exhaustively uses all regions from the domain, sometimes in practice one may decide to use only part of $Dom(X)$ to fit the parametric model. In this case only data falling in this interested part of $Dom(X)$ is used and the nonnull partition Condition 5 is imposed on this selected part of $Dom(X)$, instead of on the entire domain.

Remark 7 (Condition on mean estimates).

In regard to Condition 3, if we use $\hat{E}_{KM}(Y | X \in R_k) = \int_0^{Z(n_k)} \{1 - \hat{F}_k(u)\} du$ as described in Section 2, then according to Gill (1983), under some conditions described below,

$$\sqrt{n_k}(\hat{E}_{KM}(Y | X \in R_k) - E(Y | X \in R_k)) \xrightarrow{d} N(0, \sigma_k)$$

as $n_k \rightarrow \infty$ for $k = 1, \dots, m$, conditional on a partition $\mathcal{A}$. The expression of σ_k , which is the asymptotic variance of the Kaplan-Meier mean estimator, is given in Gill (1983) Corollary 3.2.

According to Gill (1983), the following second-level conditions suffice for the asymptotic normality of the Kaplan-Meier mean estimates (Condition 3) to hold.

(IC') In each region (i.e. given $X \in R_k$), C and Y are independent, $k = 1, \dots, m$.

(Tail Condition) In each region, $(1 - G) \geq c(1 - F)^\beta$ close to τ_F , for some constants $c > 0$ and $\beta < 1$, where $\tau_F = \sup\{t : F(t) < 1\}$. That is, the censoring distribution is heavier in the tail than the distribution of Y .

(Boundedness) $\bar{\mu}(t)/(1 - F(t))$ is bounded, where $\bar{\mu}(t) = \int_t^\infty \{1 - F(s)\}ds$.
This condition is equivalent to: $E(Y - t|Y \geq t)$ is bounded for $\forall t$.

Remark 8 (Secondary condition on censoring). The censoring condition **(IC')** is a secondary condition used for Condition 3. This is often a milder assumption than the censoring conditions used in other works.

Notice that condition **(IC')** is the same as the usual definition of independent censoring **(IC)** (C and Y are independent given X) if X is discrete, can only take m values, and R_k is the k th value, $k = 1, \dots, m$. In general, assuming the usual independent censoring condition **(IC)**, condition **(IC')** will hold if the following condition also holds:

(YC) In each region, either the censoring time C has the same distribution or the survival time Y has the same distribution, i.e. either C or Y is independent of X in each region.

The condition **(YC)** is approximately satisfied, if one can make either Y or C to have similar distributions at all X -locations in each region, possibly by choosing relatively small regions, or possibly by implementing a tree-kind algorithm, similar to CART or survival forest, to partition the regions of X .

Stute (1999) assumes that Y and C are independent *globally*, while **(IC')** only requires independence in each region. Mutual dependence on X might generate a dependence between Y and C globally, even when Y and C are independent conditional on X . On the other hand, it is often still reasonable to assume **(IC')**, when these regions are so chosen that the distributions of either Y or C are approximately the same at all X -locations inside each region. Our censoring assumption is even more relaxed than what is needed in the stratified versions of the KSV method, which requires the censoring time C be independent of X in each strata. Our condition **(IC')** would be satisfied if *either* Y *or* C is independent of X in each strata.

B. Proof of the Theorems

We will describe the main idea of the proof below.

These theorems are first proved *conditional* on the partition $\mathcal{A}$ of the indexes of the random observations into the m -regions, assuming $\mathcal{A}$ is such that the following event $\mathcal{E}$ holds: $n_k/n \rightarrow \pi_k > 0$ for all $k = 1, \dots, m$, as $n \rightarrow \infty$. But such event happens almost surely, due to the strong law of

large numbers ($\pi_k = P(X \in R_k) > 0$ due to Condition 5). Then an application of Lebesgue's dominated convergence theorem shows that the theorems hold unconditional on $\mathcal{A}$ also. [For example, the Lebesgue's dominated convergence theorem is applied to the conditional cdf (conditional on $\mathcal{A}$) of $\sqrt{n}(\tilde{\beta} - \beta)$ (or S_n , or T_n), the convergence of which to a normal cdf for almost all $\mathcal{A}$ implies the unconditional convergence in distribution also.]

Denote $\delta_k(b) = E(Y|X \in R_k) - E(g(X; b)|X \in R_k)$, $\tilde{\delta}_k(b) = \hat{E}_{KM}(Y|X \in R_k) - E(g(X; b)|X \in R_k)$, $\hat{\delta}_k(b) = \hat{E}_{KM}(Y|X \in R_k) - n_k^{-1} \sum_{i: X_i \in R_k} g(X_i, b)$, $Q(b) = \sum_1^m w_k \delta_k(b)^2$, $\tilde{Q}(b) = \sum_1^m w_k \tilde{\delta}_k(b)^2$, $\hat{Q}(b) = \sum_1^m w_k \hat{\delta}_k(b)^2$. Then $\tilde{\beta} = \arg \min_b \tilde{Q}(b)$, $\hat{\beta} = \arg \min_b \hat{Q}(b)$ are minimizers described before in the main text; the true parameter satisfies $\beta = \arg \min_b Q(b)$, since $Q(\beta)$ takes the least possible value 0 under the mean model (M).

The proof of the asymptotic distributional properties follows the spirit of Amemiya (1985, Theorem 4.1.3), which is based on a linear approximation to the score equation $\partial_b \tilde{Q}(\tilde{\beta}) = 0$ (or $\partial_b \hat{Q}(\hat{\beta}) = 0$) near β , which has the form $\sqrt{n}(\tilde{\beta} - \beta) = -\{\partial_b^2 Q(\beta)\}^{-1} \sqrt{n} \partial_b \tilde{Q}(\beta) + o_p(1)$ (or $\sqrt{n}(\hat{\beta} - \beta) = -\{\partial_b^2 Q(\beta)\}^{-1} \sqrt{n} \partial_b \hat{Q}(\beta) + o_p(1)$).

Then the theorems are proved by using the Slutsky's theorem and using the property that $\partial_b \tilde{Q}(\beta)$ is asymptotically zero-mean normal (AN0) [$\partial_b \hat{Q}(\beta)$ is a sum of two AN0 quantities].

The above described approach has used these critical assumptions. (i). It assumes that the minimizers satisfies the zero derivative conditions $\partial_b \tilde{Q}(\tilde{\beta}) = 0$ (or $\partial_b \hat{Q}(\hat{\beta}) = 0$). (ii). It assumes that $\tilde{\beta}$ and $\hat{\beta}$ are consistent for β so a linear approximation around β has small enough remainder term. (iii). It assumes that the second order derivatives of $\tilde{Q}$ or $\hat{Q}$, evaluated at some intermediate points between $\tilde{\beta}$ and β (or $\hat{\beta}$ and β) are consistent for $\partial_b^2 Q(\beta)$, which should be positive definite. (iv). It also assumes the asymptotic normal (AN) conditions of the first derivatives [$\sqrt{n} \partial_b \tilde{Q}(\beta)$ is AN0, $\sqrt{n} \partial_b \hat{Q}(\beta)$ is a sum of two AN0's]. These assumptions will be validated below, conditional on a random partition $\mathcal{A}$ such that event $\mathcal{E}$ holds, as described in the beginning of the proof.

Part 1. AN conditions (iv) (as $n \rightarrow \infty$):

The AN conditions are easiest to be proved conditionally, since we will see that $\sqrt{n} \partial_b \tilde{Q}(\beta)$ is built up linearly from the $\sqrt{n_k} \tilde{\delta}_k(\beta)$'s, which are independent AN0 quantities conditional on $\mathcal{A}$, leading to the AN0 of $\tilde{Q}(\beta)$. [We cannot conclude that $\sqrt{n_k} \tilde{\delta}_k(\beta)$'s are independent *unconditionally*, since they

depend on (random) n_k 's which are constrained to sum to n .]

Note that $\sqrt{n}\partial_b\tilde{Q}(\beta) = 2 \sum_{k=1}^m \left\{ \sqrt{\frac{n}{n_k}} w_k [-\partial_b E(g(X; \beta) | X \in R_k)] \times \sqrt{n_k} \tilde{\delta}_k(\beta) \right\}$, and $\sqrt{n_k} \tilde{\delta}_k(\beta)$ for $k = 1, \dots, m$ are *independent* AN0 quantities conditionally due to Condition 3, and due to the fact that $\tilde{\delta}_k(\beta)$'s depend on data from nonoverlapping regions which are conditionally independent. The terms n/n_k 's converge to π_k^{-1} conditional on any $\mathcal{A}$ satisfying event $\mathcal{E}$. Then an application of Slutsky theorem shows that $\sqrt{n}\partial_b\tilde{Q}(\beta)$ is AN0 as $n \rightarrow \infty$, conditionally for all $\mathcal{A}$ satisfying event $\mathcal{E}$.

Similarly, $\sqrt{n}\partial_b\hat{Q}(\beta) = 2 \sum_{k=1}^m \left\{ \sqrt{\frac{n}{n_k}} w_k \left[-\partial_b n_k^{-1} \sum_{i: X_i \in R_k} g(X_i, \beta) \right] \times \sqrt{n_k} \hat{\delta}_k(\beta) \right\}$. Note that $\sqrt{n_k} \hat{\delta}_k(\beta)$ can be decomposed into two parts:

$$\sqrt{n_k} \tilde{\delta}_k(\beta) - \sqrt{n_k} \{ n_k^{-1} \sum_{i: X_i \in R_k} g(X_i, \beta) - E(g(X; \beta) | X \in R_k) \},$$

which produces the two AN0 quantities in $\sqrt{n}\partial_b\hat{Q}(\beta)$, conditionally for all $\mathcal{A}$ satisfying event $\mathcal{E}$.

These prove the AN conditions.

Part 2. Consistency condition (ii) and zero-derivative condition (i):

It is straightforward to show that Q is the uniform probability limit of both the criterion functions $\hat{Q}(b) = \sum_1^m w_k \hat{\delta}_k(b)^2$ and $\tilde{Q}(b) = \sum_1^m w_k \tilde{\delta}_k(b)^2$, over any compact set Θ of b , i.e., $\sup_{b \in \Theta} |\hat{Q}(b) - Q(b)| \rightarrow^P 0$ and $\sup_{b \in \Theta} |\tilde{Q}(b) - Q(b)| \rightarrow^P 0$ as $n \rightarrow \infty$. This is proved by showing the uniform in-probability convergence of all the $\tilde{\delta}_k()$ and $\hat{\delta}_k()$'s, due to Condition 3 and due to the law of large numbers applied on $n_k^{-1} \sum_{i: X_i \in R_k} g(X_i, b)$.

Note that $\partial_b^2 Q(b)$ is continuous due to Condition 2 and boundedness of R_k 's (due to Condition 1). Also, $\partial_b^2 Q(b)$ is positive definite at $b = \beta$, due to Condition 4, mean model (M), and positive weights w_k 's. These ensure that $Q()$ has a unique local minimum in some compact neighborhood Θ of the true parameter β , where Q takes the least possible value 0. Then the uniform in-probability convergence results above implies that the minimizer of $\hat{Q}$ and the minimizer of $\tilde{Q}$ are both converging in probability to the minimizer of Q , over the compact set Θ . This shows the consistency of $\hat{\beta}$ and $\tilde{\beta}$.

It is noted that β is in the interior of the compact neighborhood Θ ; $\hat{\beta}$ and $\tilde{\beta}$ are consistent for β , so with probability tending to one, these minimizers will eventually be inside the interior of Θ and be local minimizers having zero derivatives $\partial_b \hat{Q}(\hat{\beta}) = 0$ (or $\partial_b \tilde{Q}(\tilde{\beta}) = 0$).

Part 3. Second-derivative conditions (iii):

Similar to the proof in Part 2 regarding $\tilde{Q}()$ and $\hat{Q}()$, one can show that $\partial_{b_j}\partial_{b_k}\tilde{Q}(b)$ and $\partial_{b_j}\partial_{b_k}\hat{Q}(b)$ both converge uniformly in probability to the limit $\partial_{b_j}\partial_{b_k}Q(b)$, over any compact set Θ , for any $j, k \in \{1, \dots, q\}$. This together with the continuity of $\partial_{b_j}\partial_{b_k}Q(b)$ (implied by Condition 2 and the boundedness of R_k 's due to Condition 1), and the consistency of $\tilde{\beta}$, implies that the second order derivatives $\partial_{b_j}\partial_{b_k}\tilde{Q}(\tilde{b})$ at any intermediate values $\tilde{b}$ between $\tilde{\beta}$ and β are consistent for $\partial_{b_j}\partial_{b_k}Q(\beta)$, which form components of a positive definite matrix due to Condition 4, positive weights w_k 's, and mean model (M). Similar conclusion holds for second order derivatives of $\hat{Q}$.

These confirm all the conditions needed for the linear approximation of $\sqrt{n}(\tilde{\beta} - \beta)$ and $\sqrt{n}(\hat{\beta} - \beta)$ described before, which leads to the proof of the theorems.

Q.E.D.

C. Proof of Corollaries

Note that for any $\alpha, \eta \in (0, 1)$,

$$\begin{aligned}
 & Pr \left\{ |C^T(\hat{\beta} - \beta)| < \Phi^{-1}(1 - (1 - \eta)\frac{\alpha}{2})\sqrt{C^T \frac{V_1}{n} C} + \Phi^{-1}(1 - \eta\frac{\alpha}{2})\sqrt{C^T \frac{V_2}{n} C} \right\} \\
 = & Pr \left\{ |C^T S_n / \sqrt{n} + C^T T_n / \sqrt{n}| < \Phi^{-1}(1 - (1 - \eta)\frac{\alpha}{2})\sqrt{C^T \frac{V_1}{n} C} + \Phi^{-1}(1 - \eta\frac{\alpha}{2})\sqrt{C^T \frac{V_2}{n} C} \right\} \\
 \geq & Pr \left\{ |C^T S_n / \sqrt{n}| < \Phi^{-1}(1 - (1 - \eta)\frac{\alpha}{2})\sqrt{C^T \frac{V_1}{n} C} \cap |C^T T_n / \sqrt{n}| < \Phi^{-1}(1 - \eta\frac{\alpha}{2})\sqrt{C^T \frac{V_2}{n} C} \right\} \\
 \geq & 1 - Pr \left\{ |C^T S_n / \sqrt{n}| \geq \Phi^{-1}(1 - (1 - \eta)\frac{\alpha}{2})\sqrt{C^T \frac{V_1}{n} C} \right\} \\
 & - Pr \left\{ |C^T T_n / \sqrt{n}| \geq \Phi^{-1}(1 - \eta\frac{\alpha}{2})\sqrt{C^T \frac{V_2}{n} C} \right\} \\
 = & 1 - Pr \left\{ |C^T S_n| \geq \Phi^{-1}(1 - (1 - \eta)\frac{\alpha}{2})\sqrt{C^T V_1 C} \right\} \\
 & - Pr \left\{ |C^T T_n| \geq \Phi^{-1}(1 - \eta\frac{\alpha}{2})\sqrt{C^T V_2 C} \right\}.
 \end{aligned}$$

Take $\liminf$ on both sides. Since $S_n \xrightarrow{d} N(0, V_1)$ and $T_n \xrightarrow{d} N(0, V_2)$ as

$n \rightarrow \infty$,

$$\begin{aligned} & \lim_{n \rightarrow \infty} \inf Pr \left\{ |C^T(\hat{\beta} - \beta)| < \Phi^{-1}(1 - (1 - \eta)\frac{\alpha}{2}) \sqrt{C^T \frac{V_1}{n} C} + \Phi^{-1}(1 - \eta\frac{\alpha}{2}) \sqrt{C^T \frac{V_2}{n} C} \right\} \\ & \geq 1 - 2((1 - \eta)\frac{\alpha}{2}) - 2\eta\frac{\alpha}{2} \\ & = 1 - \alpha \end{aligned}$$

This proves Corollary 1. Corollary 2 then follows by converting a confidence interval to be a test.

Q.E.D.

REFERENCES

- Akritis, M. G. (1996), "On the Use of Nonparametric Regression Techniques for Fitting Parametric Regression Models," *Biometrics*, 52, 1342-1362.
- Akritis, M. G. (2000), "The Central Limit Theorem Under Censoring," *Bernoulli*, 6, 1109-1120.
- Amemiya, T. (1985), *Advanced Econometrics*, Cambridge, Massachusetts: Harvard University Press.
- Breiman, L. (2002), "Software for the Masses," *Third part of the WALD lectures, presented at the 277th meeting of the Institute of Mathematical Statistics, held in Banff, Alberta, Canada, July 28 to July 31, 2002.*
<http://stat-www.berkeley.edu/users/breiman/wald2002-3.pdf>
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth International Group.
- Buckley, J. and James, I. (1979), "Linear Regression with Censored Data," *Biometrika*, 66, 429-436.
- Fygenson, M. and Zhou, M. (1994), "On Using Stratification in the Analysis of Linear Regression Models with Right Censoring," *The Annals of Statistics*, 22, 747-762.
- Gill, R. (1983), "Large Sample Behaviour of the Product-Limit Estimator on the Whole Line," *The Annals of Statistics*, 11, 49-58.

- Gross, S. T. and Lai, T.L. (1996). "Nonparametric Estimation and Regression Analysis with Left Truncated and Right Censored Data," *Journal of the American Statistical Association*, 91, 1166-1180.
- Hanson, T. E. and Johnson, W. O. (2002), "Modeling regression error with mixtures of Polya trees," *Journal of the American Statistical Association* 97, 1020-1033.
- Hosmer, D. W. and Lemeshow, S. (1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, New York: Wiley.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), "Adaptive Mixtures of Local Experts," *Neural Computation*, 3, 79-87.
- Kalbfleisch, J.D. and Prentice, R.L. (2002), *The Statistical Analysis for Failure Time Data*, New York: Wiley.
- Klein, J. P. and Moeschberger, M. L. (1997), *Survival Analysis: Techniques for Censored and Truncated Data*, Springer.
- Kottas, A. and Gelfand, A. E. (2001), "Bayesian Semiparametric Median Regression Modeling," *Journal of the American Statistical Association*, 96, 1458-1468.
- Koul, H., Susarla, V., and Van Ryzin, J. (1981), "Regression Analysis with Randomly Right-Censored Data," *The Annals of Statistics*, 9, 1276-1288.
- Kuo, L. and Mallick, B. (1997), "Bayesian Semiparametric Inference for the Accelerated Failuretime Model," *Canadian Journal of Statistics*, 25, 457-472.
- Kumazawa, Y. (1987), "A Note on an Estimator of Life Expectancy with Random Censorship," *Biometrika*, 74, 655-658.
- Lehmann, E. L. (1991), *Theory of Point Estimation*, Belmont, CA: Wadsworth International Group.
- Leurgans, S. (1987), "Linear Models, Random Censoring and Synthetic Data," *Biometrika*, 74, 301-309.
- Miller, R.G. (1976). "Least Squares Regression with Censored Data," *Biometrika*, **63**, 449-464.
- Miller, R. and Halpern, J. (1982), "Regression with Censored Data," *Biometrika*, 69, 521-531.

- Stute, W. (1994). "The bias of Kaplan-Meier Integrals," *Scandinavian Journal of Statistics*, 21, 475-484.
- Stute, W. (1995), "The Central Limit Theorem Under Random Censorship," *Annals of Statistics*, 23, 422-439.
- Stute, W. (1999), "Nonlinear Censored Regression," *Statistica Sinica*, 9, 1089-1102.
- Yang, S. (1999), "Censored Median Regression Using Weighted Empirical Survival and Hazard Functions," *Journal of the American Statistical Association*, 94, 137 -145.
- Ying, Z., Jung, S. H. and Wei, L. J. (1995), "Survival Analysis with Median Regression Models," *Journal of the American Statistical Association*, 90, 178-184.
- Zhou, M. (1999), "Regression Analysis with Censored data: Extensions of Koul-Susarla-Van Ryzin Approach," *Statistics and Probability Letters*, 41, 229-236.