University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2004 *Paper* 49

Semiparametric binary regression under monotonicity constraints

Moulinath Banerjee[*] Pinaki Biswas[†]

Debashis Ghosh[‡]

[*] University of Michigan, moulib@umich.edu

[†] Univeristy of Michigan, pbiswas@umich.edu

[‡] University of Michigan, debashis.ghosh@ucdenver.edu

Semiparametric binary regression under monotonicity constraints

Moulinath Banerjee, Pinaki Biswas, and Debashis Ghosh

Abstract

Summary: We study a binary regression model where the response variable Δ is the indicator of an event of interest (for example, the incidence of cancer) and the set of covariates can be partitioned as (X,Z) where Z (real valued) is the covariate of primary interest and X (vector valued) denotes a set of control variables. For any fixed X, the conditional probability of the event of interest is assumed to be a monotonic function of Z. The effect of the control variables is captured by a regression parameter β. We show that the baseline conditional probability function (corresponding to $X=0$) can be estimated by isotonic regression procedures and develop a likelihood ratio based method for constructing confidence intervals for this function that obviates the need to estimate nuisance parameters from the data. We also show how confidence intervals for the regression parameter can be constructed using asymptotically $\chî2$ likelihood ratio statistics. The confidence sets for the regression parameter and those for the conditional probability function are combined using Bonferroni's inequality to construct conservative confidence intervals for the conditional probability of the event of interest at different fixed values of X and Z. We present simulation results to illustrate the theory and apply our results to a prostate cancer data set.

Semiparametric Binary Regression Under Monotonicity Constraints

Moulinath Banerjee [1], Pinaki Biswas and Debashis Ghosh

University of Michigan

November 14, 2004

Abstract

We study a binary regression model where the response variable Δ is the indicator of an event of interest (for example, the incidence of cancer) and the set of covariates can be partitioned as (X, Z) where Z (real valued) is the covariate of primary interest and X (vector valued) denotes a set of control variables. For any fixed X, the conditional probability of the event of interest is assumed to be a monotonic function of Z. The effect of the control variables is captured by a regression parameter β. We show that the baseline conditional probability function (corresponding to $X = 0$) can be estimated by isotonic regression procedures and develop a likelihood ratio based method for constructing confidence intervals for this function that obviates the need to estimate nuisance parameters from the data. We also show how confidence intervals for the regression parameter can be constructed using asymptotically χ^2 likelihood ratio statistics. The confidence sets for the regression parameter and those for the conditional probability function are combined using Bonferroni's inequality to construct conservative confidence intervals for the conditional probability of the event of interest at different fixed values of X and Z. We present simulation results to illustrate the theory and apply our results to a prostate cancer data set.

Running Head: Binary Regression Under Monotonicity

1 INTRODUCTION

The study of shape-restricted functions arises extensively in statistical modeling. The nature of the shape-restriction is generally dictated by the underlying science, or empirical evidence and is generally useful in narrowing down the class of models that the statistician might want to consider. This, on one hand, often makes the statistical analysis more tractable and on the other actually leads to more informative results if the shape restriction is the right one imposed. Common

[1] Research supported in part by National Science Foundation grant DMS-0306235

Key words and phrases. binary regression, Brownian Motion, Cox model, current status data, greatest convex minorant, likelihood ratio statistic, nonregular problem, χ^2 distribution

examples of such shape restrictions are monotonicity, convexity or concavity, unimodality etc. Monotonicity, in particular, is a shape-restriction that shows up very naturally in the analysis of statistical data. The motivating example for this paper is a case in point. In screening studies involving cancer biomarkers, it has been empirically observed that increasing levels of a biomarker are often associated with elevated risk of cancer. For example, prostate-specific antigen (PSA) has been used for detection of prostate cancer. If a man has a PSA measurement between 4 and 10 ng/mL, then this leads to a prostate needle biopsy. Higher levels of PSA are assumed to be associated with increased disease risk. The question of scientific interest is the extent to which PSA measurements are predictive of prostate cancer.

More generally, in many other biological and scientific contexts, monotonicity is a natural assumption. One setting involves the study of height trajectories in adolescents (Ramsay and Silverman (1997)). It seems reasonable that during this period of time, children's heights are increasing with age. Another example comes from rheumatology (Bloch and Silverman (1997)). The goal in that work was to quantify the efficacy index and toxicity in two patient populations. It was assumed that toleration of drug was an increasing function of treatment effectiveness and a decreasing function of toxicity. Monotonicity has also played an important role is econometrics. In modelling labor participation as a function of wage, for example, classical theory of supply and demand implies that increased wages should be associated with increased labor force participation. Another situation comes from options pricing theory, in which the price of a call option is assumed to be a monotone decreasing function of the strike price.

Consequently, there has been a great deal of literature devoted to nonparametric estimation with monotonicity constraints. A relatively recent summary of such work is found in Robertson et.al.(1988). Since then, there has been substantial research on nonparametric isotonic regression procedures. A rate of convergence for estimators of monotone regression functions was established by Van der Geer (1990). Mammen (1991) analyzed rates of convergence for two-step estimators involving kernel smoothing and isotonic regression. Alternative methodology for producing a monotone estimate from an initial smooth was developed by Hall and Huang (2001). Hypothesis testing procedures involving monotonicity have been proposed by Ghosal, Sen and Van der Vaart (2000), Gijbels and Heckman (2000) and Hall and Heckman (2000). Procedures focusing more on the algorithmic aspects of nonparametric regression with monotonicity constraints have been given by Dykstra and Robertson (1982), Friedman and Tibshirani (1984), Villalobos and Wahba (1987), He and Shi (1998), Ramsay (1998) and Mammen et. al. (2001).

The preceding literature deals with the problem of studying nonparametric regression models with monotone constraints. In many applications, however, semiparametric modelling procedures are more useful. Advantages of this approach include concise summaries of covariate effects in the parametric component of the model. In contrast to the previous paragraph, semiparametric modelling procedures with monotonicity constraints have been less studied. Hastie and Tibshirani (1990) and Shiboski (1998) suggest combining backfitting procedures with isotonic regression algorithms for estimation in some of these settings. However, the asymptotic results concerning such procedures are not available. On the other hand, Huang (2002) derives some asymptotic results in a semiparametric model with a continuous response. However, no algorithmic characterizations are presented in that article. In addition, generalizations to noncontinuous outcomes have not been well–addressed in the literature. Regarding problems of inference involving semiparametric models, Murphy and Van der Vaart (1997) have noted that inference based on Wald-type statistics tends to be fairly unstable and have developed an elegant theory of semiparametric likelihood ratio inference. However, their focus is on the finite-dimensional component in the model and not on the infinite-dimensional component. In fact, in a later paper (Murphy and Van der Vaart (2000)), inference for the parametric component of the model is made by profiling out the infinite dimensional component (which is treated as a nuisance parameter) and the profile likelihood function is shown to have some of the crucial properties of usual parametric likelihoods. However, this approach does not suffice when interest focuses on both the parametric and nonparametric components of the model.

Binary regression models are used frequently to model the effects of covariates on dichotomous outcome variables. The most well known of these methods is logistic regression. In parametric logistic regression, the log-odds of observing an event is modelled as a linear function of the covariates. More generally, parametric binary regression models can be formulated as follows: If Δ is the indicator of the outcome and X is a set of covariates believed to influence the outcome, one can write

$$g(\mu(X,Z)) = \beta^T X \tag{1.1}$$

where $\mu(X,Z) = P(\Delta = 1 \mid X, Z)$ and g can be taken to be a smooth monotone increasing function from $(0,1)$ to $(-\infty,\infty)$ and is called the "link function". Models of this kind are fairly well–studied in the literature on generalized linear models (see, for example, McCullagh and Nelder (1989)) and algorithms for computing the vector of regression coefficients are well–known. Commonly used link functions are the logit (logistic regression), the probit and the complementary log-log (cloglog). Our interest is in situations where in addition to X, there is an additional covariate Z whose effect on the outcome variable is known qualitatively. More specifically, it is known that higher values of Z are associated with higher chances of an outcome ($\Delta = 1$). To incorporate the effect of Z in the model, and to ensure the monotonicity (monotone increasing) of the conditional probability of an outcome in Z, we extend (1.1) in the following way. We write,

$$g(\mu(X,Z)) = \beta^T X + \xi(Z)\,, \tag{1.2}$$

where $\xi(z)$ is some monotone function of Z. Thus the nonparametric component affects the conditional probability of a positive outcome additively on the scale of the link function. Also note that this implies that $\mu(X,Z)$ is monotone increasing in Z for every fixed X.

Models of this kind are useful in a variety of settings and have been studied by various authors. Dunson (2004) considers nonparametric estimation of ξ as in (1.2) above from a Bayesian angle. More generally, binary regression models where the conditional mean of the outcome variable is monotone in one of the regressors have been studied in econometric contexts by Magnac and Maurin (2003) (see also Manski and Tamer (2002)). Freitag (2004) considers interval estimation for quantiles in monotone binary regression models with applications to current status data. Our interest in models of this sort was triggered by some data involving biomarker and other covariate measurements on a group of individuals being monitored for prostate cancer, obtained from the CARET study (see Etzioni et. al. (1999)). In this example, the outcome variable Δ is the indicator of prostate cancer, while Z is an appropriate transform of free and bound PSA in the serum and higher values of Z are believed to be associated with greater risk for cancer. Control covariates (like age) also need to be adjusted for; we discuss these issues in greater detail in Section 5.

In this article, we consider the use of likelihood ratio inference-based methods in a semiparametric binary regression model of the type described in (1.2). For the sake of concreteness we focus on a particular link function – the cloglog link. Thus, (1.2) reduces to:

$$\log\left(-\log(1-\mu(X))\right) = \beta^T X + \xi(Z)\,,$$

where ξ is a monotone increasing function diverging to $-\infty$ as the argument converges to 0 and diverging to ∞ as the argument diverges to ∞. The covariate Z, without loss of generality, can be assumed to lie in a compact interval contained in the positive axis.

A question that may naturally arise is the use of the particular cloglog link. There are two main reasons for this. Firstly, the use of the cloglog link ensures that the resulting likelihood function for the data is suitably concave in both the finite dimensional and infinite dimensional parameter. This makes the computation of MLE's tractable and allows convenient characterizations of these. However, other link functions (like the logit) also enjoy this property. Secondly, as will be seen in Section 2, under our proposed modelling scheme, the likelihood for the data is identical to the likelihood under the Cox Proportional Hazards Model with Current Status Data. This is a problem that has been fairly well studied in the recent past (see, for example Huang (1994), Huang (1996), Murphy and Van der Vaart (1997)); hence many of the techniques and results from this model can be fairly easily adapted to our setting. Writing $\mu(Z,X) = g(\beta, Z, X)$ and $\Lambda(\beta, Z, X) = -\log\left(1 - g(\beta, Z, X)\right)$ and $\Lambda(Z) = \exp(\xi(Z))$, the above model can easily be written as:

$$\Lambda(\beta, Z, X) = \exp(\beta^T X)\,\Lambda(Z)\,.$$

Setting $\Lambda(Z) = -\log\left(1 - g(Z)\right)$, we can write:

$$P(\Delta = 0 \mid Z, X) = 1 - g(\beta, Z, X) = (1 - g(Z))^{\exp(\beta^T X)}\,,$$

where $g(\cdot)$ is an increasing and continuously differentiable function defined on $[0, \infty)$ with $g(0) = 0$ and $\lim_{z\to\infty} g(z) = 1$. For each fixed X, $g(\beta, \cdot, X)$ can be thought of as a distribution function on the positive half–line. We call $g(\beta, \cdot, 0) \equiv g(\cdot)$ the baseline conditional probability function. The function $\Lambda(Z)$ is the cumulative hazard function corresponding to $g(Z)$ whereas $\Lambda(\beta, Z, X)$ is the cumulative hazard function corresponding to $g(\beta, Z, X)$.

The likelihood for a single observation and connections to the Proportional Hazards Model: The density function of the vector (Δ, Z, X) can be written as:

$$p_{\beta,\Lambda}(\delta, z, x) = (1 - \exp(-\Lambda(z)\exp(\beta^T x)))^{\delta}\,(\exp(-\Lambda(z)\exp(\beta^T x)))^{1-\delta}\,f(z, x)\,, \tag{1.3}$$

where $f(z,x)$ is the joint density of (Z,X) with respect to Leb $\times\, \mu$ where Leb denotes Lebesgue measure on $[0,\infty)$ and μ is some measure defined on $\mathbb{R}^d$ where d is the dimension of X. But the above joint density (1.3) is identical to that for the Cox Proportional Hazards Model with current status data. To see this consider the following scenario: Let T denote the survival time of an individual, Y denote the time they are observed at and W denote a vector of covariate measurements on the individual. Suppose that T and Y are conditionally independent given W. Further suppose that one only observes $(D = 1(T \leq Y), Y, W)$. Let $\Lambda(t \mid w)$ denote the conditional hazard function of the survival time T given $W = w$. Suppose that

$$\Lambda(t \mid w) = \Lambda(t) \exp(\beta^T w)\,.$$

This is the Cox Proportional Hazards assumption with Λ acting as the baseline hazard. Suppose that the joint distribution of (Y,W) is described by the density function $f(\cdot,\cdot)$. We can now write down the joint density of (D,Y,W) quite easily. Using the conditional independence of T and Y given W, we find that the conditional distribution of D given $(Y,W) = (y,w)$ is Bernoulli $(F(y \mid w))$. Here, $F(\cdot \mid w)$ is the conditional distribution of T given $W = w$. Thus, the conditional density of D given $(Y = y, W = w)$ is

$$p(d \mid y, w) = F(y \mid w)^d \, (1 - F(y \mid w))^{1-d}\,.$$

Substituting

$$F(y \mid w) = 1 - \exp(-\Lambda(y \mid w)) = 1 - \exp(-\Lambda(y) \exp(\beta^T w))$$

into the previous display, we obtain the joint density of (D,T,W) as,

$$\tilde{p}(d,y,w) = (1 - \exp(-\Lambda(y)\exp(\beta^T w)))^{\delta} \, (\exp(-\Lambda(y)\exp(\beta^T w)))^{1-\delta} \, f(y,w)\,. \tag{1.4}$$

Comparing (1.4) with (1.3), we find that they are identical. This implies that the joint distribution of (I,Y,W) is the same as the joint distribution of (Δ, Z, X). Thus the sample $\{\Delta_i, Z_i, X_i\}_{i=1}^n$ at hand may be regarded as a sample from the Cox PH model with current status censoring (with Δ_i denoting the current status of the i'th observation, Z_i denoting the observation time and X_i the vector of control covariates).

Our main focus in this paper will be to make inferences on β, the regression parameter and g (equivalently Λ), the baseline conditional probability function using likelihood ratios. This will involve studying the likelihood ratio statistic for the following testing problems: (a) $H_0 : \beta = \beta_0$ and (b) $\tilde{H}_0 : \Lambda(z_0) = \theta_0$ for some fixed point z_0 in the domain of Z. Note that (b) is equivalent to testing for the value of g at a particular point. While inferences for β and g can be carried out using the limit distributions of the corresponding maximum likelihood estimates, we do not adopt this route, because the corresponding limit distributions involve nuisance parameters that can be difficult to estimate. On the other hand, the likelihood ratio statistics, as will be shown, are asymptotically pivotal quantities with fixed and known limit distributions and confidence intervals may be readily constructed by inverting the acceptance region of the likelihood ratio tests with thresholds determined by the quantiles of the limiting pivotal distributions. The superiority of likelihood ratio based confidence intervals over Wald type ones (which the limit distribution theory for the MLE's would yield) is well known; see the discussion in the introduction of Murphy and Van der Vaart (1997) and Chapter 1 of Banerjee (2000).

While the likelihood ratio statistic for testing $H_0 : \beta = \beta_0$ can be studied by applying the theory of Murphy and Van der Vaart (1997), the likelihood ratio procedure for testing the value of Λ at a fixed point (or multiple points) which we deal with in this paper has hitherto never been studied. We will show that the likelihood ratio statistic for testing $\tilde{H}_0 : \Lambda(z_0) = \theta_0$ converges in distribution to the random variable $\mathbb{D}$, which is a very well characterized functional of standard two–sided Brownian motion with parabolic drift. It can be thought of as an analogue of the χ_1^2 distribution (from a likelihood ratio perspective) in nonregular statistical problems involving $n^{1/3}$ rate of convergence for maximum likelihood estimators and non–Gaussian limit distributions. Indeed the maximum likelihood estimator $\hat{\Lambda}_n$ converges to the true Λ at rate $n^{1/3}$ in this problem, despite $\sqrt{n}$ rate of convergence for $\hat{\beta}$.

Our new result is a powerful one – it gives a simple and yet elegant way of estimating Λ (equivalently g) without having to estimate limiting quantiles. Our result is equally applicable to the problem of estimating the baseline survival function in the Cox PH model with current status data (because of the correspondence between this model and ours, as illustrated above).

The rest of the paper is organized as follows. Maximum likelihood estimation and novel likelihood ratio-based inferential procedures are discussed in Section 2. The associated asymptotic results, which are also new, are given in Section 3. The finite-sample properties of the proposed methods are assessed using simulation studies and with application to data from a prostate cancer study in Section 4. We conclude with some discussion in Section 5. Proofs of some of the results in Section 4 are collected in the Appendix (Section 6).

2 COMPUTING MLE'S AND LIKELIHOOD RATIOS

In what follows, we denote the true underlying values of the parameters (β, Λ) by (β_0, Λ_0). The log–likelihood function for the sample, up to an additive factor that does not involve any of the parameters of interest, is given by,

$$l_n(\beta, \Lambda) = \sum_{i=1}^{n} \left[\Delta_i \, \log\left(1 - \exp(-\Lambda(Z_i)\exp(\beta^T X_i))\right) - (1 - \Delta_i)\exp(\beta^T X_i)\,\Lambda(Z_i)\right] .$$

Let $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}$ denote the ordered values of the Z_i's; let $\Delta_{(i)}$ and $X_{(i)}$ denote the indicator and covariate values associated with biomarker value $Z_{(i)}$. Also, let $\Lambda_i \equiv \Lambda(Z_{(i)})$ and $R_i(\beta) = \exp(\beta^T X_{(i)})$. For $\delta \in \{0, 1\}$ and $r, u \geq 0$ set,

$$\phi(\delta, r, u) = -\delta \, \log(1 - e^{-r\,u}) + (1 - \delta)\, r\, u \,. \tag{2.5}$$

It is easy to check that ϕ is convex in u; also,

$$-l_n(\beta, \Lambda) \equiv \Psi(\beta, \Lambda) = \sum_{i=1}^{n} \phi(\Delta_{(i)}, R_i(\beta), \Lambda_i) \,.$$

The parameter set for β is taken to be a bounded subset $\mathcal{C} \subset \mathbb{R}^d$. Here d is the dimension of the covariate X.

Minimizing Ψ with respect to β and Λ amounts to finding

$$\left(\hat{\beta}_n, (\hat{\Lambda}_{n,1}, \hat{\Lambda}_{n,2}, \ldots, \hat{\Lambda}_{n,n})\right) = \text{argmin}_{\beta \in \mathcal{C}\ ,\ 0 \le u_1 \le u_2 \le \ldots \le u_n} \sum_{i=1}^{n} \phi(\Delta_{(i)}, R_i(\beta), u_i)\,.$$

Thus the MLE of Λ is only identifiable up to its values at the $Z_{(i)}$'s. This does not cause a problem as far as the asymptotic results are concerned; however, for the sake of concreteness we take $\hat{\Lambda}_n$, the MLE of Λ to be the (unique) right–continuous increasing step function that assumes the value $\hat{\Lambda}_{n,i}$ at the point $Z_{(i)}$ and has no jump points outside of the set $\{Z_{(i)}\}_{i=1}^n$.

Let $\hat{\Lambda}_n^{(\beta)} = \text{argmin}_\Lambda \Psi(\beta, \Lambda)$. As above, we can compute $\hat{\Lambda}_n^{(\beta)}$ uniquely only up to its values at the $Z_{(i)}$'s and indeed, we identify it with this vector. Thus,

$$\hat{\beta}_n = \text{argmin}_\beta \Psi(\beta, \hat{\Lambda}_n^{(\beta)}) \ \text{ and } \ \hat{\Lambda}_n = \hat{\Lambda}_n^{(\hat{\beta}_n)}\,.$$

The likelihood ratio statistic for testing $H_0 : \beta = \beta_0$ is given by:

$$\text{lrtbeta}_n = 2\,(l_n(\hat{\beta}_n, \hat{\Lambda}_n) - l_n(\beta_0, \hat{\Lambda}_n^{(\beta_0)}))\,. \tag{2.6}$$

We next discuss the computation of the constrained maximizers of β and Λ, say $(\hat{\beta}_{n,0}, \hat{\Lambda}_{n,0})$ under $\tilde{H}_0 : \Lambda(z_0) = \theta_0$ with $0 < \theta_0 < \infty$. As in the unconstrained case, this maximization can be achieved in two steps. For each β, one can compute

$$\hat{\Lambda}_{n,0}^{(\beta)} = \text{argmin}_{\Lambda : \Lambda(z_0) = \theta_0} \Psi(\beta, \Lambda)\,.$$

Then,

$$\hat{\beta}_{n,0} = \text{argmin}_\beta \Psi(\beta, \hat{\Lambda}_{n,0}^{(\beta)}) \ \text{ and } \ \hat{\Lambda}_{n,0} = \hat{\Lambda}_{n,0}^{(\hat{\beta}_{n,0})}\,.$$

The likelihood ratio statistic for testing $\tilde{H}_0 : \Lambda(z_0) = \theta_0$ is given by:

$$\text{lrtg}_n = 2\,(l_n(\hat{\beta}_n, \hat{\Lambda}_n) - l_n(\hat{\beta}_{n,0}, \hat{\Lambda}_{n,0}))\,. \tag{2.7}$$

One way of computing the MLE's of β and Λ in practice is to vary β on a sufficiently fine grid over its domain, compute $\Psi(\beta, \hat{\Lambda}_n^{(\beta)})$ for each β on the grid and select that value on the grid for which this quantity is minimized. This is in fact what Huang (1996) does. The MLE's of β and Λ under $\tilde{H}_0$ can be computed similarly. The main disadvantages of the grid search procedure are computational intensity (especially in higher dimensions) and the discretization bias. The latter can of course be reduced by refining the grid but only at the expense of increased computational intensity. While we did implement the grid search procedure in our simulation studies, some alternative methods of computing β were also investigated. While there is a heuristic aspect to the alternative procedures, we found them to work extremely well on simulated data sets, producing results in conformity with those produced by grid–search. The main advantage of these alternative methods is that they are much faster.

The alternative methods are based on $\dot{l}_{n,\beta}(\beta,\Lambda) \equiv (\partial/\partial\,\beta)\;\; l_n(\beta,\Lambda)$, the score function for β. We have

$$\frac{\partial}{\partial\,\beta}\,\Psi(\beta,\Lambda) = -\dot{l}_{n,\beta}(\beta,\Lambda) = -\sum_{i=1}^{n}\left(\Delta_{(i)}\,\frac{\exp(-\Lambda(Z_{(i)})\,R_i(\beta))}{1-\exp(-\Lambda(Z_{(i)})\,R_i(\beta))} - (1-\Delta_{(i)})\right)\times\Lambda(Z_{(i)})\,R_i(\beta)\,X_{(i)}\,.$$

Now, $(\hat{\beta}_n, \hat{\Lambda}_n)$ clearly solve

$$\frac{\partial}{\partial\,\beta}\,\Psi(\beta,\Lambda) = 0\,. \tag{2.8}$$

However, this is not the unique solution. If we define $\hat{\beta}_n(\Lambda)$ to be the minimizer of $\Psi(\beta,\Lambda)$ for a fixed Λ, then clearly $(\hat{\beta}_n(\Lambda),\Lambda)$ satisfies (2.8). However, one can try to find a zero of (2.8) in the set $\{(\beta,\hat{\Lambda}_n^{(\beta)}) : \beta \in \mathcal{C})\}$. Since $\hat{\beta}_n(\hat{\Lambda}_n^{(\beta)})$ is not guaranteed to be equal to β, a pair of the type $(\beta,\hat{\Lambda}_n(\beta))$ will not satsify (2.8) in general. However, $\hat{\beta}_n(\hat{\Lambda}_n) = \hat{\beta}_n$, so we are guaranteed at least one solution, namely the MLE's of β and Λ. Though we were not able to establish that any root of (2.8) of the form $(\beta,\hat{\Lambda}_n^{(\beta)})$ must necessarily be the MLE, this did turn out to be the case for fairly extensive simulation studies. We solve

$$\frac{\partial}{\partial\,\beta}\,\Psi(\beta,\hat{\Lambda}_n^{(\beta)}) = 0$$

in the following manner.

(0) Choose an intial value $\beta^{(0)}$ and a small number ϵ.

(1) Set $\beta = \beta^{(0)}$. Compute $\hat{\Lambda}_n^{(\beta)}$.

(2) Solve,

$$\frac{\partial}{\partial\,\gamma}\,\Psi(\gamma,\hat{\Lambda}_n^{(\beta)}) = 0\,.$$

Set $\beta^{(0)}$ to be equal to the solution. If $\mid\, \beta^{(0)} - \beta \,\mid < \epsilon$, stop. Otherwise go to Step (1).

The above method can be adapted in a straightforward manner for computing the MLE's under $\tilde{H}_0$. We omit a discussion. The method proposed above is similar to that in Zhang (2002) for computing maximum likelihood estimates in a semiparametric model involving panel count data.

We focus now on the computation of $\hat{\Lambda}_n^{(\beta)}$ and $\hat{\Lambda}_{n,0}^{(\beta)}$.

Characterizing $\hat{\Lambda}_n^{(\beta)}$: This is characterized by the vector $0 \le \hat{\Lambda}_{n,1}^{(\beta)} \le \ldots \le \hat{\Lambda}_{n,n}^{(\beta)}$ that minimizes the expression,

$$\psi\,(\beta,u) = \sum_{i=1}^{n}\,\phi(\Delta_{(i)}, R_i(\beta), u_i)$$

over all $0 \leq u_1 \leq u_2 \leq \ldots \leq u_n$. Without loss of generality one can assume that $\Delta_{(1)} = 1$ and $\Delta_{(n)} = 0$. If not, the effective sample size for the estimation of the parameters is $k_2 - k_1 + 1$ where k_1 is the first index i such that $\Delta_{(i)} = 1$ and k_2 is the last index such that $\Delta_{(i)} = 0$. It is not difficult to see that one can set $\hat{\Lambda}_{n,i}^{(\beta)} = 0$ for all $i < k_1$ and $\hat{\Lambda}_{n,i}^{(\beta)} = \infty$ for all $i > k_2$ without imposing any constraints on the other components of the mininimizing vector.

The function $\psi(\beta, u)$ which for brevity we will denote by ψ can be minimized using standard methods from convex optimization theory. Using the Kuhn–Tucker theorem for minimizing a convex function subject to linear constraints, we obtain a set of necessary and sufficient conditions (*Fenchel conditions*) which are as follows:

$$\sum_{j=i}^{n} \frac{\partial\, \phi(\Delta_{(j)}, R_j(\beta), u_j)}{\partial\, u_j}\, (\hat{u}_j) \geq 0 \quad \text{for} \quad i = 1, 2, \ldots, n \tag{2.9}$$

and

$$\sum_{j=1}^{n} \hat{u_j}\, \frac{\partial\, \phi(\Delta_{(j)}, R_j(\beta), u_j)}{\partial\, u_j}\, (\hat{u}_j) = 0\,. \tag{2.10}$$

Let $B_1, B_2, \ldots, B_k$ be the blocks of indices on which the solution $\hat{u}$ is constant (these are called level blocks) and let w_i be the common value on block B_i. Under our assumption that $\Delta_{(1)} > 0$ it must be the case that $w_1 > 0$. Then, on each B_i, we have that

$$\sum_{j \in B_i} \frac{\partial\, \phi(\Delta_{(j)}, R_j(\beta), u_j)}{\partial\, u_j}\, (w_i) = 0\,.$$

Thus w_i is the unique solution to the equation

$$\sum_{j \in B_i} \frac{\partial\, \phi(\Delta_{(j)}, R_j(\beta), u_j)}{\partial\, u_j}\, (w) = 0\,.$$

The solution $\hat{u}$ can be viewed as the slope of the greatest convex minorant (slogcm) of a cumulative sum diagram. This characterization is needed for the asymptotic theory. The basic idea is to use the Fenchel conditions above to formulate a quadratic optimization problem under monotonicity constraints whose solution still remains $\hat{u}$ and then appeal to standard results from the theory of isotonic regression. Details of this procedure can be found in Banerjee (2004). We omit the details here but provide the "self–consistency" characterization of $\hat{u}$. For $1 \leq i \leq n$, set $d_i = \bigtriangledown_{ii}\, \psi(\hat{u})$. Define the function ξ as follows:

$$\begin{aligned} \xi(u) &= \sum_{i=1}^{n} \left[u_i - \hat{u}_i + \bigtriangledown_i\, \psi(\hat{u})\, d_i^{-1}\right]^2\, d_i \\ &= \sum_{i=1}^{n} \left[u_i - \left(\hat{u}_i - \bigtriangledown_i\, \psi(\hat{u})\, d_i^{-1}\right)\right]^2\, d_i\,. \end{aligned}$$

It can be shown that $\hat{u}$ minimizes ξ subject to the constraints that $0 \leq u_1 \leq u_2 \leq \ldots \leq u_n$ and hence furnishes the isotonic regression of the function

$$g(i) = \hat{u}_i - \bigtriangledown_i \psi(\hat{u})\, d_i^{-1}$$

on the ordered set $\{1, 2, \ldots, n\}$ with weight function $d_i \equiv \bigtriangledown_{ii} \psi(\hat{u})$. It is well known that the solution

$$(\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n) = \text{slogcm} \left\{ \sum_{j=1}^{i} d_i \, , \, \sum_{j=1}^{i} g(i)\, d_i \right\}_{i=0}^{n} .$$

See, for example Theorem 1.2.1 of Robertson et.al.(1988).

Since $\hat{u}$ is unknown, we need to iterate. Thus, we pick an initial guess for $\hat{u}$, say $u^{(0)}$ and satisfying the monotonicity constraints, compute $u^{(1)}$ by solving the isotonic regression problem discussed above, plug in $u^{(1)}$ as an updated guess for $\hat{u}$, obtain $u^{(2)}$ and proceed thus, until convergence. However there are convergence issues with a simple minded iterative scheme of the above type, since the algorithm could hit inadmissible regions in the search space. Jongbloed (1998) addresses this issue by using a modified iterated convex minorant (MICM) algorithm; see Section 2.4 for a discussion of the practical issues and a description of the relevant algorithm which incorporates a line search procedure to guarantee convergence to the desired value. We provide explicit forms for the points d_i and g_i in the current situation. We have

$$d_i = \frac{\partial^2}{\partial\, u_i^2}\, \psi(u) = \frac{\Delta_{(i)}\, R_i(\beta)^2\, e^{-R_i(\beta)\, u_i}}{(1 - e^{-R_i(\beta)\, u_i})^2}$$

and

$$g(i) = u_i - \bigtriangledown_i \psi(u)\, d_i^{-1}$$

, with

$$\bigtriangledown_i \psi(u) = -\frac{\Delta_{(i)}\, e^{-R_i(\beta)\, u_i}\, R_i(\beta)}{1 - e^{-R_i(\beta)\, u_i}} + (1 - \Delta_{(i)})\, R_i(\beta)\, .$$

The algorithm stops when the Fenchel conditions (2.9) and (2.10) are satisfied to a pre-specified degree of tolerance.

An important consequence of the above self–consistency characterization is the fact that on each block (of indices) B_i where $\hat{u}$ is constant, the common solution can be written as a weighted average of the g_j's for the j's in that block, with the weights given by the d_j's. We now introduce some notation that will prove useful later. Denote $\phi(\Delta_{(i)}, R_i(\beta), t)$ by $\phi_{i,\beta}(t)$ and its first and second derivatives with respect to t by $\phi'_{i,\beta}(t)$ and $\phi''_{i,\beta}(t)$. Then we can write

$$\hat{\Lambda}_n^{(\beta)} \equiv \text{slogcm} \left\{ \sum_{i=1}^{k} \phi''_{i,\beta}(\hat{\Lambda}_n^{(\beta)}(Z_{(i)}))\, , \, \sum_{i=1}^{k} \left[\hat{\Lambda}_n^{(\beta)}(Z_{(i)}) - \frac{\phi'_{i,\beta}(\hat{\Lambda}_n^{(\beta)}(Z_{(i)}))}{\phi''_{i,\beta}(\hat{\Lambda}_n^{\beta}(Z_{(i)}))} \right] \phi''_{i,\beta}(\hat{\Lambda}_n^{(\beta)}(Z_{(i)})) \right\}_{k=0}^{n} .$$

Hence, we can write w_i, the common value of the solution $\hat{\Lambda}_n^{(\beta)}$ on the block B_i, as

$$\hat{\Lambda}_n^{(\beta)}(Z_{(j)}) = \frac{\sum_{k \in B_i} \{\hat{\Lambda}_n^{(\beta)}(Z_{(k)})\, \phi''_{k,\beta}(\hat{\Lambda}_n^{(\beta)}(Z_{(k)})) - \phi'_{k,\beta}(\hat{\Lambda}_n^{(\beta)}(Z_{(k)}))\}}{\sum_{k \in B_i} \phi''_{k,\beta}(\hat{\Lambda}_n^{(\beta)}(Z_{(k)}))} \quad \text{for } \; j \in B_i\,. \tag{2.11}$$

Characterizing $\hat{\Lambda}_{n,0}^{(\beta)}$: Let m be the number of biomarker values that are less than or equal to z_0. Finding $\hat{\Lambda}_{n,0}^{(\beta)}$ amounts to minimizing

$$\psi(\beta, u) = \sum_{i=1}^{n} \phi(\Delta_{(i)}, R_i(\beta), u_i)$$

over all $0 \leq u_1 \leq u_2 \ldots \leq u_m \leq \theta_0 \leq u_{m+1} \leq \ldots \leq u_n$. This can be reduced to solving two separate optimization problems. These are:

$$(1) \qquad \text{Minimize} \sum_{i=1}^{m} \phi(\Delta_{(i)}, R_i(\beta), u_i) \text{ over } 0 \leq u_1 \leq u_2 \leq \ldots \leq u_m \leq \theta_0.$$

$$(2) \qquad \text{Minimize} \sum_{i=m+1}^{n} \phi(\Delta_{(i)}, R_i(\beta), u_i) \text{ over } \theta_0 \leq u_{m+1} \leq u_{m+2} \leq \ldots \leq u_n.$$

Consider (1) first. As in the unconstrained minimization problem one can write down the Kuhn–Tucker conditions characterizing the minimizer. It is then easy to see that the solution $(\hat{u}_1^{(0)}, \hat{u}_2^{(0)}, \ldots, \hat{u}_m^{(0)})$ can be obtained through the following recipe. Minimize $\sum_{i=1}^{m} \phi(\Delta_{(i)}, R_i(\beta), u_i)$ over $0 \leq u_1 \leq u_2 \leq \ldots \leq u_m$ to get $(\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_m)$. Then,

$$(\hat{u}_1^{(0)}, \hat{u}_2^{(0)}, \ldots, \hat{u}_m^{(0)}) = (\tilde{u}_1 \wedge \theta_0, \tilde{u}_2 \wedge \theta_0, \ldots, \tilde{u}_m \wedge \theta_0)\,.$$

The solution vector to (2), say $(\hat{u}_{m+1}^{(0)}, \hat{u}_{m+2}^{(0)}, \ldots, \hat{u}_n^{(0)})$ is similarly given by

$$(\hat{u}_{m+1}^{(0)}, \hat{u}_{m+2}^{(0)}, \ldots, \hat{u}_n^{(0)}) = (\tilde{u}_{m+1} \vee \theta_0, \tilde{u}_{m+2} \vee \theta_0, \ldots, \tilde{u}_n \vee \theta_0)\,,$$

where

$$(\tilde{u}_{m+1}, \tilde{u}_{m+2}, \ldots, \tilde{u}_n) = \text{argmin}_{u_{m+1} \leq u_{m+2} \leq \ldots \leq u_n} \sum_{i=m+1}^{n} \phi(\Delta_{(i)}, R_i(\beta), u_i)\,.$$

A careful examination of the relationship of the unconstrained solution to the constrained solution reveals that:

$$\hat{\Lambda}_n^{(\beta)}(z) \neq \hat{\Lambda}_{n,0}^{(\beta)}(z) \Rightarrow \hat{\Lambda}_{n,0}^{(\beta)}(z_0) = \theta_0 \;\text{ or }\; \hat{\Lambda}_n^{(\beta)}(z) = \hat{\Lambda}_n^{(\beta)}(z_0)\,. \tag{2.12}$$

The constrained solution also has a "self–consistent" characterization in terms of the slope of the greatest convex minorant of a cumulative sum diagram. This follows in the same way as for the unconstrained solution by using the Kuhn–Tucker theorem and formulating a quadratic optimization problem based on the Fenchel conditions given by this theorem. We skip the details but give the self-consistent characterization.

The constrained solution $\hat{u}^{(0)}$ minimizes,

$$A(u_1, u_2, \ldots, u_n) = \sum_{i=1}^{n} \left[u_i - \left(\hat{u}_i^{(0)} - \bigtriangledown_i \, \psi(\hat{u}^{(0)}) d_i^{-1} \right) \right]^2 d_i$$

subject to the constraints that $0 \le u_1 \le u_2 \le \ldots \le u_m \le \theta_0 \le u_{m+1} \le \ldots \le u_n$ and hence furnishes the isotonic regression of the function

$$g(i) = \hat{u}_i^{(0)} - \bigtriangledown_i \, \psi(\hat{u}^{(0)}) \, d_i^{-1}$$

on the ordered set $\{1, 2, \ldots, n\}$ with weight function $d_i \equiv \bigtriangledown_{ii} \, \psi(\hat{u}^{(0)})$. Here $\psi \equiv \psi(\beta, u)$ as before. The constrained solution can be found, as in the unconstrained case, by using the MICM. An important consequence of the "self–consistent" characterization is that on each block $\tilde{B}$ of indices on which $\hat{u}^{(0)}$ is constant and not equal to θ_0, it can be written as $\sum_{i \in \tilde{B}} g(i) \, d_i \; / \; \sum_{i \in \tilde{B}} d_i$. Let $\tilde{B}_1, \tilde{B}_2, \ldots, \tilde{B}_p$ denote the blocks of indices on which $\hat{u}^{(0)}$ is constant and let $\{\tilde{w}_i\}_{i=1}^{p}$ denote the corresponding set of values. Thus, as long as $\tilde{w}_i \neq \theta_0$, it can be written as

$$\tilde{w}_i \equiv \hat{\Lambda}_{n,0}^{(\beta)}(Z_{(j)}) = \frac{\sum_{k \in \tilde{B}_i} \{ \hat{\Lambda}_{n,0}^{(\beta)}(Z_{(k)}) \, \phi''_{k,\beta}(\hat{\Lambda}_{n,0}^{(\beta)}(Z_{(k)})) - \phi'_{k,\beta}(\hat{\Lambda}_{n,0}^{(\beta)}(Z_{(k)})) \}}{\sum_{k \in \tilde{B}_i} \phi''_{k,\beta}(\hat{\Lambda}_{n,0}^{(\beta)}(Z_{(k)}))} \quad \text{for } \; j \in \tilde{B}_i \, . \qquad (2.13)$$

This representation will prove useful later on.

3 ASYMPTOTIC RESULTS

In this section we present asymptotic results for the estimation of β and g. The parameter space for β is taken to be an open bounded subset of $\mathbb{R}^d$. We denote it by $\mathcal{C}$. The parameter space for Λ is the space of all nondecreasing cadlag (i.e., right-continuous with left-hand limits) functions from $[0, \tau]$ to $[0, M]$ where M is some large positive constant. Let (β_0, Λ_0) denote the true model parameters. We make the following assumptions:

(A.1) The true regression parameter β_0 is an interior point of $\mathcal{C}$.

(A.2) The covariate X has bounded support. Hence, there exists x_0 such that $P(\|X\| \le x_0) = 1$. Also $E(\mathrm{Var}(X \mid Z))$ is positive definite with probability one.

(A.3) Let g_0 denote the true baseline conditional probability function. Then $g_0(0) = 0$. Let $\tau_{g_0} = \inf\{z : g_0(z) = 1\}$. The support of Z is an interval $[\sigma, \tau]$ with $0 < \sigma < \tau < \tau_{g_0}$.

Remarks: The boundedness of $\mathcal{C}$ along with assumptions (A.1)–(A.3) are imposed to deduce the consistency and rates of convergence of the maximum likelihood estimators (see page 546 of Huang (1996)) of β and Λ. In particular, the boundedness of the covariate X does not cause a problem with applications. The utility of the assumption that the conditional dispersion of X given Z is poistive definite is explained below. Further assumptions follow; of these (A.4) and (A.5) are fairly weak regularity conditions on the true baseline conditional probability function and the distribution of the biomarker. The assumption (A.6) is a very technical assumption and is require to ensure that one can define appropriate approximately *least favorable submodels* as in Murphy and Van der Vaart (1997) (pages 1483–1484). These are crucial for deriving the limit distribution of the likelihood ratio statistic for testing for the regression parameter.

(A.4) Let $\Lambda_0 = -\log(1-g_0)$. We assume that $0 < \Lambda_0(\sigma-) < \Lambda(\tau) < M$. Also, Λ_0 is continuously differentiable on $[\sigma,\tau]$ with derivative λ_0 bounded away from 0 (and automatically from ∞).

(A.5) The marginal density of Z is continuous and positive on $[\sigma,\tau]$.

(A.6) The function $h^{\star\star}$ given by (3.14) has a version which is differentiable componentwise with each component possessing a bounded derivative on $[\sigma,\tau]$.

We now introduce the efficient score function for β in this model. Recall that the joint density of the vector (Δ, Z, X) is given by:

$$p_{\beta,\Lambda}(\delta,z,x) = (1-\exp(-\Lambda(z)\exp(\beta^T x)))^{\delta}\,(\exp(-\Lambda(z)\exp(\beta^T x)))^{1-\delta}\,f(z,x)\,.$$

The ordinary score function for β in this model is:

$$\dot{l}_\beta(\beta,\Lambda)(\delta,z,x) = (\partial/\partial\,\beta)\,\log\,p_{\beta,\Lambda}(\delta,x,z) = x\,\Lambda(z)\,Q((\delta,z,x);\theta,\Lambda)\,,$$

where

$$Q((\delta,z,x);\theta,\Lambda) = e^{\beta^T x}\,\left[\delta\,\frac{\exp(-e^{\beta^T x}\,\Lambda(z))}{1-\exp(-e^{\beta^T x}\,\Lambda(z))} - (1-\delta)\right]\,.$$

The score function for Λ is a linear operator acting on the space of functions of bounded variation on $[\sigma,\tau]$ and has the form:

$$\dot{l}_\Lambda(\beta,\Lambda)(h(\cdot))(\delta,z,x) = h(z)\,Q((\delta,z,x);\theta,\Lambda)\,.$$

Here h is a function of bounded variation on $[\sigma,\tau]$. The efficient score function for β at the true parameter values (β_0,Λ_0), which we will denote by $\tilde{l}$ for brevity, is defined as

$$\tilde{l} = \dot{l}_\beta(\beta_0,\Lambda_0) - \dot{l}_\Lambda(\beta_0,\Lambda_0)h^\star$$

for functions $h^\star = (h_1^\star, h_2^\star, \ldots, h_d^\star)$ of bounded variation, such that $h_i^\star$ minimizes the distance

$$E_{\beta_0,\Lambda_0}(\dot{l}_{\beta,i}(\beta_0,\Lambda_0) - \dot{l}_\Lambda(\beta_0,\Lambda_0)\,h(\cdot))^2\,,$$

for h varying in the space of functions of bounded variation on $[\sigma,\tau]$. Here

$$\dot{l}_{\beta,i}(\beta_0,\Lambda_0) = x_i\,\Lambda(z)\,Q((\delta,z,x);\beta_0,\Lambda_0)$$

is the i'th component of the ordinary score function for β. The problem of finding $h_i^\star$ for each i is a weighted least squares problem and the solution to $h^\star$ can be easily seen to be given by:

$$h^\star(Z) = \Lambda_0(Z)\, h^{\star\star}(Z) = \Lambda_0(Z)\, \frac{E_{\beta_0,\Lambda_0}(Z\,Q^2((\Delta,Z,X);\beta_0,\Lambda_0 \mid Z))}{E_{\beta_0,\Lambda_0}(Q^2((\Delta,Z,X);\beta_0,\Lambda_0 \mid Z))}\,. \tag{3.14}$$

The assumption that $E(\mathrm{Var}(X \mid Z))$ is positive definite (A.2) ensures that $\tilde{l}$ the efficient score function for β is not identically zero, whence the efficient information $\tilde{I}_0 = \mathrm{Disp}(\tilde{l}) \equiv E_{\beta_0,\Lambda_0}(\tilde{l}\,\tilde{l}^T)$ is positive definite (Note that $E_{\beta_0,\Lambda_0}(\tilde{l}) = 0$). This entails that the MLE of β will converge at $\sqrt{n}$ rate to the true value and have an asymptotically normal distribution with a finite dispersion matrix.

Now consider the problem of testing $H_0 : \beta = \beta_0$ based on our data, but under the (true) constraint that $\Lambda(z_0) = \theta_0$. Thus, we define:

$$\mathrm{lrtbeta}_n^0 = 2\,\log\,\frac{\mathrm{argmax}_{\Lambda(z_0)=\theta_0}\, l_n(\beta,\Lambda)}{\mathrm{argmax}_{\beta=\beta_0,\Lambda(z_0)=\theta_0}\, l_n(\beta,\Lambda)}\,. \tag{3.15}$$

Thus,

$$\mathrm{lrtbeta}_n^0 = 2\,l_n(\hat{\beta}_{n,0},\hat{\Lambda}_{n,0}) - 2\,l_n(\beta_0,\hat{\Lambda}_{n,0}^{(\beta_0)})\,.$$

We now state a theorem describing the asymptotic behavior of $\hat{\beta}_n$ and $\hat{\beta}_{n,0}$ (which we subsequently denote by $\tilde{\beta}_n$) and the likelihood ratio statistics $\mathrm{lrtbeta}_n$ as defined in (2.6) and $\mathrm{lrtbeta}_n^0$ above.

Theorem 3.1 *Under Conditions (A.1) – (A.7), both $\hat{\beta}_n$ and $\tilde{\beta}_n$ are asymptotically linear in the efficient score function and have the following representation:*

$$\sqrt{n}\,(\hat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}}\,\tilde{I}_0^{-1}\,\sum_{i=1}^n \tilde{l}(\Delta_i, Z_i, X_i) + r_n$$

and

$$\sqrt{n}\,(\tilde{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}}\,\tilde{I}_0^{-1}\,\sum_{i=1}^n \tilde{l}(\Delta_i, Z_i, X_i) + s_n$$

where r_n and s_n are $o_p(1)$. Hence both $\sqrt{n}\,(\hat{\beta}_n - \beta_0)$ and $\sqrt{n}\,(\tilde{\beta}_n - \beta_0)$ converge in distribution to $N(0,\tilde{I}_0^{-1})$.

Furthermore,

$$\mathit{lrtbeta}_n = n(\hat{\beta}_n - \beta_0)^T\,\tilde{I}_0\,(\hat{\beta}_n - \beta_0) + o_p(1)\,, \tag{3.16}$$

while

$$\mathit{lrtbeta}_n^0 = n(\tilde{\beta}_n - \beta_0)^T\,\tilde{I}_0\,(\tilde{\beta}_n - \beta_0) + o_p(1)\,. \tag{3.17}$$

It follows that both $\mathit{lrtbeta}_n$ and $\mathit{lrtbeta}_n^0$ are asympotically distributed like χ_d^2.

We do not provide a detailed proof of this theorem in this paper. The properties of $\hat{\beta}_n$ and $\mathrm{lrtbeta}_n$ stated in the theorem can be deduced by arguments similar to those in Theorem 3.4 of Huang (1996) and the treatment of the Cox Proportional Hazards Model with Current Status Data in Murphy and Van der Vaart (1997). The derivation in Murphy and Van der Vaart (1997) is done for a one–dimensional β but the proof extends easily to higher dimensions. An alternative route to the asymptotic distribution of $\hat{\beta}_n$ is to adapt the arguments in Section A.3 of Murphy and Van der Vaart (1997). The asymptotically linear representation for $\tilde{\beta}_n$ and the limiting χ^2 distribution for $\mathrm{lrtbeta}_n^0$ follows in analogous fashion. Some additional care needs to be exercised, since the parameter space for Λ is now restricted by fixing the value at the point z_0. Roughly the intuition is the following: $\hat{\beta}_n$, the unconstrained MLE of β, is $\sqrt{n}$–consistent and asymptotically efficient for the given model. The unconstrained likelihood ratio statistic for testing $\beta = \beta_0$, which we denote by $\mathrm{lrtbeta}_n$, is asymptotically χ^2. These properties will be preserved even when we compute the above statistics under the single (true) constraint that $\Lambda(z_0) = \theta_0$. In fact, the same asymptotic representations for the above statistics will continue to hold when we constrain Λ at finitely many points. Note however, that the limit distribution of the MLE will generally be affected under infinitely many constraints on Λ. This is easily seen when we constrain Λ on the support of Z. In this case Λ is completely known and the asymptotic variance of β is the inverse of the ordinary information for θ as opposed to the efficient information.

We next state asymptotic results concerning the nonparametric component of the model. In order to do so, we introduce the following processes. For positive constants c and d define the process $X_{c,d}(z) := c\,W(z) + d\,z^2$, where $W(z)$ is standard two-sided Brownian motion starting from 0. Let $G_{c,d}(z)$ denote the GCM (greatest convex minorant) of $X_{c,d}(z)$. Then $g_{c,d}(z)$ is the right derivative of $G_{c,d}$ and can be shown to be a piecewise constant (increasing) function, with finitely many jumps in any compact interval. Next, let $G_{c,d,L}(h)$ denote the GCM of $X_{c,d}(h)$ on the set $h \leq 0$ and $g_{c,d,L}(h)$ denote its right–derivative process. For $h > 0$, let $G_{c,d,R}(h)$ denote the GCM of $X_{c,d}(h)$ on the set $h > 0$ and $g_{c,d,R}(h)$ denote its right–derivative process. Define $g^0_{c,d}(h)$ as $g_{c,d,L}(h) \wedge 0$ for $h \leq 0$ and as $g_{c,d,R}(h) \vee 0$ for $h > 0$. Then $g^0_{c,d}(h)$, like $g_{c,d}(h)$, is a piecewise constant (increasing) function, with finitely many jumps in any compact interval and differing (almost surely) from $g_{c,d}(h)$ on a finite interval containing 0. In fact, with probability 1, $g^0_{c,d}(h)$ is identically 0 in some (random) neighborhood of 0, whereas $g_{c,d}(h)$ is almost surely non-zero in some (random) neighborhood of 0. Also, the interval $D_{c,d}$ on which $g_{c,d}$ and $g^0_{c,d}$ differ is $O_p(1)$. For more detailed descriptions of the processes $g_{c,d}$ and $g^0_{c,d}$, see Banerjee (2000), Banerjee and Wellner (2001) and Wellner (2001). Thus, $g_{1,1}$ and $g^0_{1,1}$ are the unconstrained and constrained versions of the slope processes associated with the canonical process $X_{1,1}(z)$. By Brownian scaling, the slope processes $g_{c,d}$ and $g^0_{c,d}$ can be related in distribution to the canonical slope processes $g_{1,1}$ and $g^0_{1,1}$. This is the content of the following proposition.

Lemma 3.1 *For any $M > 0$, the following distributional equality holds in the space $L_2[-M, M] \times L_2[-M, M]$:*

$$\left(g_{c,d}(h), g^0_{c,d}(h)\right) \overset{\mathcal{D}}{=} \left(c\,(d/c)^{1/3} g_{1,1}\left((d/c)^{2/3} h\right), c\,(d/c)^{1/3} g^0_{1,1}\left((d/c)^{2/3} h\right)\right).$$

Here $L_2[-M, M]$ denotes the space of real–valued functions on $[-M, M]$ with finite L_2 norm (with respect to Lebesgue measure).

This is proved in BANERJEE (2000), Chapter 3.

Let z_0 be an interior point of the support of Z. Now, define the (localized) slope processes U_n and V_n as follows:

$$U_n(h) = n^{1/3}\,(\hat{\Lambda}_n^{(\beta_0)}(z_0 + h\,n^{-1/3}) - \Lambda_0(z_0)) \text{ and } V_n(h) = n^{1/3}\,(\hat{\Lambda}_{n,0}^{(\beta_0)}(z_0 + h\,n^{-1/3}) - \Lambda_0(z_0))\,.$$

The following theorem describes the limiting distribution of the slope processes above.

Theorem 3.2 *Define,*

$$C(z_0) = \int \frac{e^{2\,\beta_0^T x}\, exp(-e^{\beta_0^T x}\,\Lambda_0(z_0))}{1 - exp(-e^{\beta_0^T x}\,\Lambda_0(z_0))}\, f(z_0, x)\, d\,\mu(x)\,.$$

Assume that $0 < C(z_0) < \infty$. Let

$$a = \sqrt{\frac{1}{C(z_0)}} \text{ and } b = \frac{1}{2}\,\lambda_0(z_0)\,,$$

where λ_0 is the derivative of Λ_0. The processes $(U_n(h), V_n(h))$ converge finite dimensionally to the processes $(g_{a,b}(h), g_{a,b}^0(h))$. Furthermore, using the monotonicity of the processes U_n and V_n, it follows that the convergence holds in the space $L_2[-K, K] \times L_2[-K, K]$ for any $K > 0$.

We now describe the limiting behavior of the likelihood ratio statistic for testing (the true) $\tilde{H}_0 : \Lambda(z_0) = \theta_0$.

Theorem 3.3 *The likelihood ratio statistic for testing $\tilde{H}_0 : \Lambda(z_0) = \theta_0$ as defined in (2.7) converges in distribution to $\mathbb{D}$ where*

$$\mathbb{D} = \int \left((g_{1,1}(z))^2 - (g_{1,1}^0(z))^2\right)\, dz\,.$$

We end this section with the statement and discussion of a conjecture which has been supported very well by simulation experiments.

Conjecture: The likelihood ratio statistic for testing $\mathcal{H}_0 : \beta = \beta_0, \Lambda(z_0) = \theta_0$ converges in distribution to the convolution of $\mathbb{D}$ and χ_d^2.

Denote the likelihood ratio statistic for testing $\mathcal{H}_0$ by L_n. The conjecture is motivated by the fact that

$$L_n = \text{lrtbeta}_n + 2\,(l_n(\beta_0, \hat{\Lambda}_n^{(\beta_0)}) - l_n(\beta_0, \hat{\Lambda}_{n,0}^{(\beta_0)}))\,.$$

The first term on the right side of the above display is asymptotically distributed like χ_d^2 and the second term, by the proof of Theorem 3.3, is distributed asymptotically like $\mathbb{D}$. Thus, for the conjecture to hold, we need the asymptotic independence of the first and the second term. This fact still remains to be established; however, numerical simulations strongly support the above result.

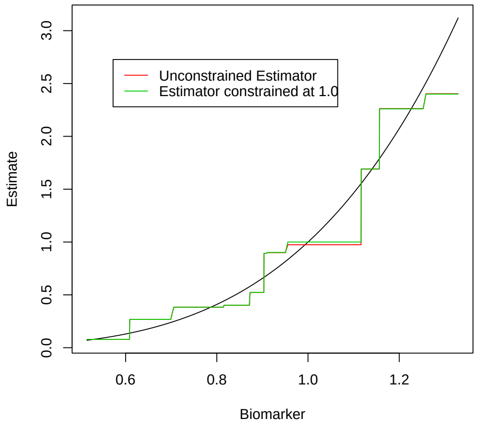

Figure 1: Constrained and unconstrained nonparametric estimators of $\Lambda_0(t)$.

4 NUMERICAL COMPARISONS

In this section we report results from simulation studies and data analysis.

4.1 SIMULATION STUDIES

In an attempt to assess the performance of the techniques described above, extensive simulation studies were conducted, of which only a fraction are reported here.

Data were generated as $\{(\Delta_i, Z_i, X_i) : i = 1, 2, \ldots, n\}$ from $p_{\beta,\Lambda}(\cdot)$, as defined in Section 1, for fixed values of the parameters β and $\Lambda(\cdot)$. For simplicity, X_i was assumed to be univariate.

Two choices were considered for the joint distribution of (Z_i, X_i). In the first, we assumed independence of Z_i and X_i with X_i being normally distributed with mean 0 and variance 1 truncated to lie in [-2, 2], whereas Z_i has an Exponential distribution with mean 1 and truncated to [0.5, 1.5]. In the second case, X_i has the same distribution as in the first case but conditional on X_i, Z_i has a truncated Exponential distribution on [0.5, 1.5] with mean $8/(8 + X_i)$. The conditional distribution of Δ_i is thus Bernoulli with probability $p_i = 1 - \exp\{-\Lambda(Z_i)e^{\beta X_i}\}$. This set–up can be verified to satisfy assumptions (A.1) – (A.6) of Section 3.

Sample sizes $n = 200, 500$ and 1000 were considered. We took $\beta = -0.5, -0.25, 0.0, 0.25$ and 0.5. The choice for $\Lambda(t)$ was taken as $\Lambda(t) = t^4$. Simulations for a linear and concave shaped $\Lambda(t)$ were also carried out; the estimators performed similarly to that reported here. Likelihood ratio tests of hypotheses of the form $H_0 : \beta = \beta_0$ and $\tilde{H}_0 : \Lambda(z_0) = \theta_0$ were carried out resulting in confidence intervals for β and $\Lambda(z_0)$. For a fixed value of x_0, a confidence interval for $p_0 = 1 - \exp\{-\Lambda(z_0)e^{\beta x_0}\}$, was also constructed using the confidence intervals for β and $\Lambda(z_0)$ simultaneously, using the Bonferroni procedure to adjust for the overall coverage probability. The target coverage for the individual intervals for β and $\Lambda(z_0)$ was taken as 97.5% so that one expects to have a coverage of at least 95% for p_0. Coverage probabilities and average lengths were estimated from 1000 replications.

Tables 1 and 2 demonstrate the performance of the confidence intervals, where for each fixed value of β and a combination of (z_0, x_0), the first row refers to the coverage probability, with the second row showing the average length of the interval. As expected, the intervals get narrower with increasing sample size. The coverage for $\Lambda(z_0)$ is not affected by changes in β or (z_0, x_0), although the coverage for β gets affected as it moves farther from zero. The coverage for p_0 turns out to be conservative in almost all of the cases.

Figure 1 demonstrates the performance of the unconstrained and constrained nonparametric estimators of $\Lambda(t)$ for a single sample of size 1000 and $z_0 = 1.0$. The estimators are doing quite well, and differing only in a neighborhood of $z_0 = 1.0$, as expected. Figure 2 displays the quantile-quantile plot for the likelihood ratio statistic for testing $\tilde{H}_0$ and it can be seen that the statistic is performing well. Finally, Table 3 demonstrates the fact that the likelihood ratio statistic approaches the limit earlier than $\sqrt{n}(\hat{\beta} - \beta)$, the centered and scaled MLE $\hat{\beta}$, in finite samples. This is carried out by regressing the quantiles of the likelihood ratio statistic on that of the χ_1^2 distribution and by regressing the quantiles of $\sqrt{n}(\hat{\beta} - \beta)$ on that of the $N(0,1)$ distribution. There is a significant intercept in the latter case, indicative of a bias, whereas the intercept and slope from the former case reflect the true values more closely. The true value of β in Table 3 was chosen as 0.25.

4.2 PROSTATE CANCER DATA

The procedures in the paper were motivated by and implemented on a prostate cancer data set obtained from the CARET study (for the details, see Etzioni et. al. (1999)). Prostate specific antigen (PSA) measured in serum is currently used as a biomarker for prostate cancer. Free and bound levels of PSA were measured in 71 subjects who developed prostate cancer ($\Delta = 1$) and 71 age-matched controls ($\Delta = 0$), all of whom participated in the study. The 71 prostate cancer cases were diagnosed between September 1988 and September 1995. Each case was assigned a matched control, namely, a study subject who had not been diagnosed as having prostate cancer by the time of analysis. Subjects who participated had serum drawn and stored at entry into the study and at 2 year intervals thereafter. The inspection times for PSA were measured with reference to the time of diagnosis. For the purpose of our analyses, we considered only the most recent PSA measurement relative to diagnosis for each subject. We considered $Z = -\log$(free PSA / total PSA) as the biomarker. A typical assumption in biomarker studies is that increasing levels of the biomarker is associated with increased risk of developing the disease. Therefore, in the semiparametric model, we modelled the effect of Z on disease risk as a monotonic function. The covariates included in the parametric component of the model included age and time relative to diagnosis.

Maximum likelihood estimates for the regression parameter $\boldsymbol{\beta}$ were obtained and likelihood ratio tests were conducted for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ using the methods of Section 2. Based on the parameter estimates, the parametric component is highly significant; the p-value for $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is less than 1.1×10^{-16}). The estimated coefficient for age is numerically very small ($\hat{\beta}_2 = 0.015$), due to the fact that age had been adjusted for in the design of the study by enrolling age-matched controls. The coefficient for time relative to diagnosis is negative ($\hat{\beta}_1 = -3.562$), thereby implying that smaller time to diagnosis is associated with larger probability of prostate cancer. Figure 3 displays the joint confidence set obtained for $\boldsymbol{\beta}$ in two dimensions and Figure 4 displays the estimate of $\Lambda(z)$ as discussed in Section 2.

Confidence intervals for $\Lambda(z_0)$ and the disease probability p_0 were also obtained for some choices of z_0 and $\boldsymbol{x}_0 = (x_{01}, x_{02})'$. The choice for z_0 was taken as the median value of the biomarker at 1.46. The graph of Λ (the cumulative hazard corresponding to the baseline conditional probability function) also indicates that there is a change in the curve around the median value; to the left, the cumulative hazard is almost flat whereas it starts rising fairly quickly to the right. A number of choices were made for $\boldsymbol{x}_0$ and are summarized in Table 4. The target confidence level for p_0 was taken at 95%. It is clear from Table 4 that the disease probability increases with increasing age and decreases with increasing time relative to diagnosis. However, the data do not seem to carry much information about disease incidence for smaller values of times relative to diagnosis (the first three rows of the table are not very informative).

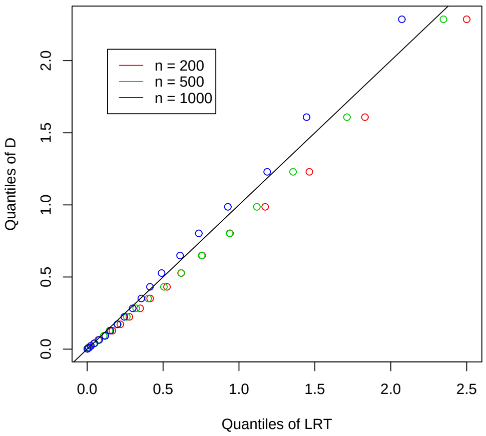

Figure 2: Performance of the Likelihood Ratio Statistic for testing $\tilde{H}_0$ when Z is independent of X.

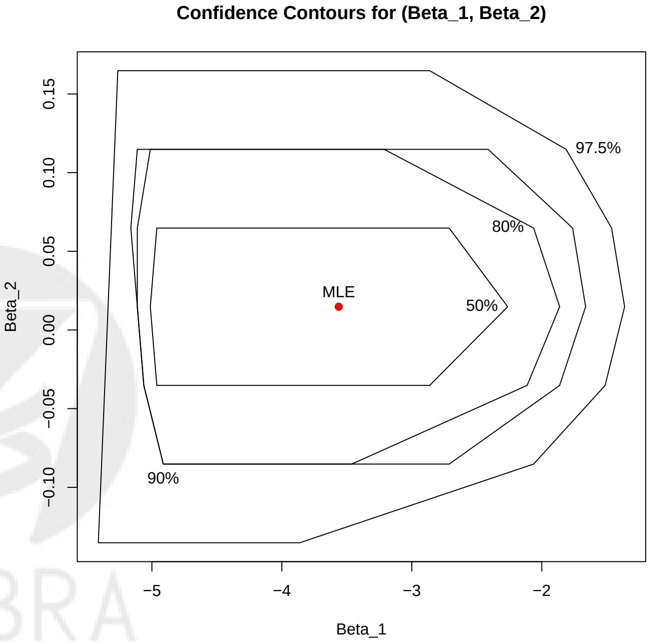

Figure 3: Confidence Set for $\boldsymbol{\beta}$.

Table 1: Table showing simulation results when Z_i and X_i are independent.

$z_0 = 0.8$ *and* $x_0 = -1.0$

	$n = 200$			$n = 500$			$n = 1000$		
β	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)
-0.5	95.2	98.2	99.7	95.9	97.8	99.8	96.5	96.9	99.8
	0.776	0.563	0.637	0.455	0.405	0.464	0.311	0.320	0.361
-0.25	96.3	97.9	99.6	95.6	97.3	100	96.3	98.4	99.9
	0.721	0.545	0.721	0.423	0.414	0.421	0.291	0.314	0.320
0.0	96.7	97.9	99.6	97.3	98.2	99.7	97.2	97.7	99.8
	0.706	0.561	0.517	0.413	0.399	0.361	0.283	0.315	0.279
0.25	95.8	97.8	99.6	95.8	97.5	99.3	97.3	96.5	99.5
	0.726	0.564	0.446	0.424	0.409	0.315	0.290	0.314	0.238
0.5	94.0	97.5	99.4	94.6	97.6	99.5	94.7	97.2	98.9
	0.781	0.551	0.373	0.456	0.403	0.265	0.312	0.314	0.203

$z_0 = 1.0$ *and* $x_0 = 1.0$

	$n = 200$			$n = 500$			$n = 1000$		
β	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)
-0.5	94.7	97.7	99.7	96.3	96.9	99.7	96.0	96.8	99.8
	0.782	1.256	0.559	0.455	0.852	0.397	0.311	0.654	0.303
-0.25	96.4	95.7	99.9	96.3	96.4	100	96.7	96.7	99.0
	0.727	1.178	0.569	0.424	0.843	0.415	0.290	0.651	0.319
0.0	97.1	97.5	99.8	97.3	96.6	99.1	97.5	97.5	99.3
	0.703	1.199	0.562	0.413	0.842	0.417	0.284	0.641	0.320
0.25	95.2	96.7	99.9	97.0	98.4	99.7	96.9	96.7	99.0
	0.726	1.224	0.529	0.424	0.860	0.400	0.290	0.651	0.312
0.5	94.6	96.8	99.5	94.9	96.3	99.1	95.5	97.4	99.9
	0.776	1.244	0.480	0.455	0.871	0.360	0.312	0.655	0.280

Table 2: Table showing simulation results when Z_i and X_i are dependent.

$z_0 = 0.8$ and $x_0 = -1.0$

	$n = 200$			$n = 500$			$n = 1000$		
β	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)
-0.5	94.3	95.9	99.7	96.0	97.3	99.8	96.8	97.5	99.5
	0.779	0.550	0.629	0.455	0.406	0.463	0.313	0.316	0.359
-0.25	95.7	97.9	99.7	96.3	97.9	99.7	96.5	97.6	99.6
	0.724	0.561	0.582	0.426	0.402	0.417	0.291	0.316	0.321
0.0	96.3	97.2	99.0	96.7	96.8	99.7	97.7	97.0	99.6
	0.709	0.556	0.514	0.413	0.403	0.363	0.284	0.314	0.279
0.25	97.4	97.7	98.9	97.6	96.6	99.1	96.7	98.1	99.6
	0.727	0.554	0.443	0.424	0.408	0.314	0.290	0.315	0.239
0.5	94.0	97.3	99.3	96.5	97.5	99.3	94.6	97.9	99.9
	0.780	0.548	0.373	0.453	0.406	0.267	0.311	0.316	0.203

$z_0 = 1.0$ and $x_0 = 1.0$

	$n = 200$			$n = 500$			$n = 1000$		
β	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)	CI(β)	CI($\Lambda(z_0)$)	CI(p_0)
-0.5	93.9	95.7	99.6	94.8	97.8	99.6	95.3	96.4	98.7
	0.775	1.272	0.563	0.456	0.871	0.400	0.312	0.662	0.305
-0.25	96.7	97.2	99.5	96.3	98.1	99.2	95.6	97.4	99.6
	0.721	1.214	0.574	0.426	0.842	0.415	0.291	0.644	0.316
0.0	97.0	97.3	99.9	96.8	96.7	99.3	97.2	97.0	99.4
	0.709	1.191	0.570	0.413	0.832	0.415	0.283	0.648	0.321
0.25	96.4	97.7	99.7	96.2	97.2	99.3	97.2	97.7	99.8
	0.729	1.206	0.538	0.425	0.845	0.398	0.291	0.651	0.312
0.5	94.9	96.9	99.7	97.6	97.1	99.8	96.1	97.0	99.6
	0.776	1.252	0.480	0.454	0.873	0.361	0.311	0.661	0.283

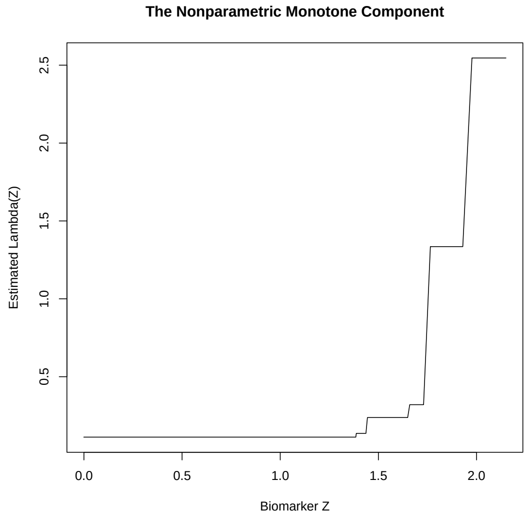

Figure 4: Estimate of $\Lambda(z)$.

5 DISCUSSION

In this paper, we have studied a semiparametric binary regression model and applied it to studying the association between biomarker levels and prostate cancer in the presence of auxiliary covariates. The effect of the auxiliary covariates is captured by a finite–dimensional regression parameter, whereas that of the biomarker is specified through a monotone increasing function. While we have used the complementary log log link in our modelling scheme our results are by no means restricted to the use of this specific link function. Link functions that preserve the concavity of the log–likelihood function in Λ and are adequately differentiable will typically work. The complementary log log link has the nice property that it relates the regression model to the Cox PH model under interval censoring.

The use of likelihood ratios for estimating both the finite and infinite dimensional components of the model proves advantageous, since nuisance parameters need no longer be estimated. Because of the natural connection to the Cox PH model, as discussed in Section 1, this work also provides a means for estimating the baseline hazard function in the Cox model under interval censoring.

Some issues remain. Firstly, the likelihood (and likelihood ratio) based approach uses step estimates of the underlying monotone function Λ. However, since the true function is smooth, it is conceivable that a smooth isotonic estimate of Λ may lead to better finite sample inference than the likelihood based method. Such smoothness constraints are typically imposed through penalized likelihood or penalized least squares criteria. This seems to be a direction for further research. Secondly, the plot of the NPMLE of Λ (Figure 4) indicates that there is a fairly marked

Table 3: Performance of the Likelihood Ratio Statistic versus the centered ML statistic.

Linear Regression of LRT(β) quantiles on χ^2_1 quantiles

Effect	Estimate	Std. Error	t value	P-Value
Intercept	-0.0002	0.0092	-0.03	0.98
Slope	1.1628	0.0069	168.26	<2e-16

Linear Regression of $\sqrt{n}(\hat{\beta} - \beta)$ quantiles on $N(0, 1)$ quantiles

Effect	Estimate	Std. Error	t value	P-Value
Intercept	0.5940	0.0122	48.54	<2e-16
Slope	2.2243	0.0142	157.03	<2e-16

Table 4: Confidence intervals for disease probability.

			Age		
Time	57.31	63.32	67.50	71.38	75.86
-0.131	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]
3.162	[0, 0.99]	[0, 1]	[0, 1]	[0, 1]	[0, 1]
3.992	[0, 0.91]	[0, 0.99]	[0, 0.99]	[0, 0.99]	[0, 1]
5.164	[0, 0.25]	[0, 0.43]	[0, 0.60]	[0, 0.76]	[0, 0.91]
5.777	[0, 0.09]	[0, 0.17]	[0, 0.26]	[0, 0.38]	[0, 0.54]

change in the behavior of the function around 1.5 (close to the median value of the biomarker). The function is reasonably flat to the left of 1.5 and takes off rapidly to its right. This suggests that instead of estimating the function pointwise, it might be of interest to estimate a threshold value for the biomarker. Determining thresholds through split– point estimation techniques have been recently studied by Banerjee and McKeague (2003) in the context of fairly general nonparametric regression problems. A semiparametric generalization of their approach to discrete outcome variables may prove fruitful for determining such a threshold value.

Finally, while the likelihood ratio method has natural advantages as illustrated in this paper, one problem with implementing it to construct confidence sets for the regression parameter β (especially in higher dimensions) is the "inversion" itself. For one dimensional β, the convexity of the log–likelihood ratio in β dictates that the confidence set be an interval and a bisection method can be resorted to. For higher dimensions however, determining the level sets of the likelihood ratio can be a tricky affair. For the data analyzed in this paper (with two–dimensional β), we used grid–search, but this is not really a feasible option in high dimensions. Apart from prohibitive computational complexity, the grid–search method only gives us a grid–based approximation to the true convex set and the possibility of obtaining better approximations to the true set through advanced computational techniques suggests itself. Such techniques, if developed fairly generically, would be useful for obtaining likelihood ratio based confidence sets in a wide variety of semiparametric problems.

6 APPENDIX

Proof–sketch of Theorem 3.2: The proof of this theorem relies on extensive use of "switching relationships" which allow us to translate the behavior of the slope of the convex minorant of a random cumulative sum diagram (this is how the estimators $\hat{\Lambda}_n^{(\beta_0)}$ and $\hat{\Lambda}_{n,0}^{(\beta_0)}$ are characterized) in terms of the minimizer of a stochastic process. The limiting behavior of the slope process can then be studied in terms of the limiting behavior of the minimizer of this stochastic process by applying argmin continuous mapping theorems. Switching relationships on the limit process then allow interpretation of the behavior of the minimizer of the limit process in terms of the slope of the convex minorant of the limiting versions of the cumulative sum diagrams (appropriately normalized).

The first step is to establish finite–dimensional convergence of the processes $(U_n(h), V_n(h))$ to $(g_{a,b}(h), g^0_{a,b}(h))$. Thus, it is shown that for any $(h_1, h_2, \ldots, h_k)$, the random vector

$$\left(\{U_n(h_i)\}_{i=1}^k, \{V_n(h_i)\}_{i=1}^k\right) \to_d \left(\{g_{a,b}(h_i)\}_{i=1}^k, \{g^0_{a,b}(h_i)\}_{i=1}^k\right) ,$$

in the space $\mathbb{R}^{2k}$. Next, to deduce the convergence in $L_2[-K, K] \times L_2[-K, K]$ note firstly that $U_n(h)$ and $V_n(h)$ are monotone functions. Now, given a sequence (ψ_n, ϕ_n) in $L_2[-K, K] \times L_2[-K, K]$ such that ψ_n and ϕ_n are monotone functions and for all vectors $(h_1, h_2, \ldots, h_k)$,

$$(\psi_n(h), \phi_n(h)) \mid_{h=h_1,h_2,\ldots,h_k} \to (\psi(h), \phi(h)) \mid_{h=h_1,h_2,\ldots,h_k}$$

where (ψ, ϕ) is in $L_2[-K, K] \times L_2[-K, K]$ it is the case that $(\psi_n, \phi_n) \to (\psi, \phi)$ in $L_2[-K, K] \times L_2[-K, K]$. It then immediately follows, in the wake of convergence of all the finite - dimensional marginals of (U_n, V_n) to those of $(g_{a,b}(h), g^0_{a,b}(h))$, that

$$(U_n(h), V_n(h)) \to_d \ (g_{a,b}(h), g^0_{a,b}(h))$$

in $L_2[-K, K] \times L_2[-K, K]$ (this parallels the result of Corollary 2 following Theorem 3 of Huang and Zhang (1994)). □

In the remainder of this proof we will sketch the proof of convergence of $U_n(h)$ to $g_{a,b}(h)$ for any h; the general proof of finite–dimensional convergence is cumbersome to write out and contains minor extensions of the ideas expounded here. In what follows, we denote $\hat{\Lambda}_n^{(\beta_0)}$ by $\tilde{\Lambda}$. For a fixed Λ we define the following processes:

$$W_{n,\Lambda}(r) = \mathbb{P}_n \left[e^{\beta_0^T X} \left(\frac{\Delta \exp(-e^{\beta_0^T X} \Lambda(Z))}{1 - \exp(-e^{\beta_0^T X} \Lambda(Z))} - (1 - \Delta) \right) 1(Z \le r) \right] ,$$

$$G_{n,\Lambda}(r) = \mathbb{P}_n \left[\Delta \, \frac{e^{2\,\beta_0^T X} \exp(-e^{\beta_0^T X} \Lambda(Z))}{(1 - \exp(-e^{\beta_0^T X} \Lambda(Z)))^2} \, 1(Z \le r) \right] ,$$

and

$$B_{n,\Lambda}(r) = W_n(r) + \int_0^r \Lambda(z) \, d\, G_{n,\Lambda}(z) \, .$$

We will denote by W_n, G_n, B_n the above processes when $\Lambda = \tilde{\Lambda}$.

We can now use "the switching relationship" for the unconstrained MLE $\tilde{\Lambda}(z)$ to get:

$$\tilde{\Lambda}(z) \le a \Leftrightarrow \operatorname{argmin}_{r \ge 0} \left[B_n(r) - a\, G_n(r)\right] \ge Z_z \tag{6.18}$$

where Z_z is the largest biomarker value not exceeding z. By argmin we denote the largest element in the set of minimizers. This can be chosen to be one of the Z_i's. The above equivalence is a direct characterization of the fact that the vector $\{\tilde{\Lambda}(Z_{(i)})\}_{i=1}^n$ is the vector of slopes (left–derivatives) of the cumulative sum diagram formed by the points $\{G_n(Z_{(i)}), B_n(Z_{(i)})\}_{i=0}^n$, computed at the points $\{G_n(Z_{(i)})\}_{i=1}^n$. The easiest way to verify this is by drawing a picture.

Now, $U_n(h_0) = n^{1/3} \, (\tilde{\Lambda}(z_0 + h_0 \, n^{-1/3}) - \Lambda_0(z_0))$. We want to find

$$\lim_{n \to \infty} P \left(n^{1/3} \, (\tilde{\Lambda}(z_0 + h_0 \, n^{-1/3}) - \Lambda_0(z_0)) \le x\right) .$$

Now, define

$$A_n = \{n^{1/3} \, (\tilde{\Lambda}(z_0 + h_0 \, n^{-1/3}) - \Lambda_0(z_0)) \le x\} \, .$$

Consider the event A_n. We have

$$\begin{aligned}
n^{1/3}\,(\tilde{\Lambda}(z_0 + h_0\, n^{-1/3}) - \Lambda_0(z_0)) \leq x \quad &\Leftrightarrow \quad \tilde{\Lambda}(z_0 + h_0\, n^{-1/3}) \leq \Lambda_0(z_0) + x\, n^{-1/3} \\
&\Leftrightarrow \quad \text{argmin}_r \left[B_n(r) - (\Lambda_0(z_0) + x\, n^{-1/3})\, G_n(r)\right] \geq Z_{(z_0 + h_0\, n^{-1/3})} \\
&\Leftrightarrow \quad \text{argmin}_r \left[V_n(r) - x\, n^{-1/3}\, G_n(r)\right] \geq Z_{(z_0 + h_0\, n^{-1/3})}\,,
\end{aligned}$$

where the second step in the above display follows from the first on using (6.18), and $V_n(r) = B_n(r) - \Lambda_0(z_0)\, G_n(r)$. Thus,

$$\begin{aligned}
A_n &= \left\{n^{1/3}\left(\text{argmin}_r \left[V_n(r) - x\, n^{-1/3}\, G_n(r)\right] - z_0\right) \geq n^{1/3}\,(Z_{(z_0 + h_0\, n^{-1/3})} - z_0)\right\} \\
&= \left\{\text{argmin}_h\, V_n(z_0 + h\, n^{-1/3}) - x\, n^{-1/3}\, G_n(z_0 + h\, n^{-1/3}) \geq h_0 + o_p(1)\right\} \\
&= \{\text{argmin}_h\, \mathbb{M}_n(h) - x\, \mathbb{G}_n(h) \geq h_0 + o_p(1)\}\ ,
\end{aligned}$$

where

$$\mathbb{M}_n(h) = n^{2/3}\left[V_n(z_0 + h\, n^{-1/3}) - V_n(z_0)\right]$$

and

$$\mathbb{G}_n(h) = n^{1/3}\left[G_n(z_0 + h\, n^{-1/3}) - G_n(z_0)\right]\ .$$

The process $\mathbb{M}_n(h) - x\, \mathbb{G}_n(h)$ converges in the space $B_{loc}(\mathbb{R})$ (here $B_{loc}(\mathbb{R})$ is the space of real–valued functions on the real line that are bounded on every compact set and equipped with the topology of uniform convergence on compact sets) to the process $L(h) \equiv \tilde{a}\, W(h) + \tilde{b}\, h^2 - x\, C(z_0)\, h$. Here $\tilde{a} = \sqrt{C(z_0)}$, $\tilde{b} = \lambda_0(t_0)\, C(z_0)/2$ and $W(h)$ is a fixed two-sided Brownian motion process starting from 0. This result is obtained by using the fact that the process $\mathbb{M}_n(h)$ converges to the limiting process $\tilde{a}\, W(h) + \tilde{b}\, h^2$ under the topology of uniform convergence on compact sets. The convergence of $\mathbb{M}_n(h)$ can be deduced from the convergence of the process

$$\tilde{P}_{n,\Lambda_0}(h) = n^{2/3}\left[B_{n,\Lambda_0}(z_0 + h\, n^{-1/3}) - B_{n,\Lambda_0}(z_0) - \Lambda_0(z_0)\,(G_{n,\Lambda_0}(z_0 + h\, n^{-1/3}) - G_{n,\Lambda_0}(z_0))\right]$$

to $\tilde{a}\, W(h) + \tilde{b}\, h^2$ (by arguments similar to those in Lemma 2.3 of Banerjee (2004)) along with the fact that $\sup_{h \in [-M,M]} \mid \tilde{\Lambda}(z_0 + h\, n^{-1/3}) - \Lambda_0(z_0) \mid = O_p(n^{-1/3})$ which entails that $\sup_{h \in [-K,K]} \mid P_{n,\Lambda_0}(h) - \mathbb{M}_n(h) \mid \to_p 0$, for every $K > 0$. Furthermore, the process $\mathbb{G}_n(h)$ converges uniformly in probability on every $[-K, K]$ to the deterministic process $C(z_0)\, h$.

The convergence in distribution of $\text{argmin}_h\, \mathbb{M}_n(h) - x\, \mathbb{G}_n(h)$ to $\text{argmin}_h\, L(h)$ is accomplished by appealing to an appropriate argmin continuous mapping theorem. The key facts that guarantee the convergence of the minimizers are (i) the fact that the limiting process possesses a unique minimizer almost surely and (ii) the minimizers of the finite sample processes are tight. This involves application of an appropriate "rate theorem" for minimizers of stochastic processes (for example Theorem 3.2.5 or Theorem 3.4.1 of Van der Vaart and Wellner (1996)). The computations are tedious but straightforward and skipped here. For a flavor of the key steps involved in establishing tightness, we refer the reader to Section 3.2.3 of Van der Vaart and Wellner (1996) and in particular Example 3.2.15 (current status data) which is naturally related to binary regression and pages 212 – 216 of Banerjee (2000).

It follows that

$$\lim_{n\to\infty} P\,(n^{1/3}\,(\tilde{\Lambda}(z_0 + h_0\,n^{-1/3}) - \Lambda_0(z_0)) \leq x) = P\,(\mathrm{argmin}_{\mathbb{R}}\,\tilde{a}\,W(h) + \tilde{b}\,h^2 - x\,C(z_0)\,h \geq h_0)\,. \tag{6.19}$$

We now use the switching relationships on the limit process. From the work of Groeneboom (1989) it follows that

$$\mathrm{argmin}_{\mathbb{R}}\,\tilde{a}\,W(h) + \tilde{b}\,h^2 - x\,C(z_0)\,h > h_0 \Leftrightarrow g_{\tilde{a},\tilde{b}}(h_0) < x\,C(z_0)\,,$$

with probability one. Therefore,

$$\lim_{n\to\infty} P\,(n^{1/3}\,(\tilde{\Lambda}(z_0 + h_0\,n^{-1/3}) - \Lambda_0(z_0)) \leq x) = P\,(g_{\tilde{a},\tilde{b}}(h_0) < x\,C(z_0))\,.$$

On noting that:

$$\frac{1}{C(z_0)}\,\left(g_{\tilde{a},\tilde{b}}(\cdot), g^0_{\tilde{a},\tilde{b}}(\cdot)\right) \equiv_d (g_{a,b}(\cdot), g^0_{a,b}(\cdot))\,,$$

with a and b as defined in the statement of the theorem, (this follows readily from Lemma 3.1) our proof is complete. □

Proof of Theorem 3.3: The likelihood ratio statistic of interest can be written as

$$\begin{aligned}
\mathrm{lrtg}_n &= 2\,(l_n(\hat{\beta}_n, \hat{\Lambda}_n) - l_n(\hat{\beta}_{n,0}, \hat{\Lambda}_{n,0})) \\
&= 2\,(l_n(\beta_0, \hat{\Lambda}_n^{(\beta_0)}) - l_n(\beta_0, \hat{\Lambda}_{n,0}^{(\beta_0)})) + 2\,(l_n(\hat{\beta}_n, \hat{\Lambda}_n) - l_n(\beta_0, \hat{\Lambda}_n^{(\beta_0)})) - 2(l_n(\hat{\beta}_{n,0}, \hat{\Lambda}_{n,0}) - l_n(\beta_0, \hat{\Lambda}_{n,0}^{(\beta_0)}))\,.
\end{aligned}$$

It will follow from Theorem 3.1 that

$$\tilde{R}_n \equiv 2\,(l_n(\hat{\beta}_n, \hat{\Lambda}_n) - l_n(\beta_0, \hat{\Lambda}_n^{(\beta_0)})) - 2(l_n(\beta_{n,0}, \hat{\Lambda}_{n,0}) - l_n(\beta_0, \hat{\Lambda}_{n,0}^{(\beta_0)}))$$

is $o_p(1)$ whence it suffices to find the asymptotic distribution of

$$C_n = 2\,(l_n(\beta_0, \hat{\Lambda}_n^{(\beta_0)}) - l_n(\beta_0, \hat{\Lambda}_{n,0}^{(\beta_0)}))\,.$$

This is precisely the likelihood ratio statistic for testing $\Lambda_0(z_0) = \theta_0$ holding β fixed at its true value β_0. We can write C_n as,

$$C_n = 2\,\left[\sum_{i=1}^{n} \phi(\Delta_{(i)}, R_i(\beta_0), \hat{\Lambda}_{n,0}^{(\beta_0)}(Z_{(i)})) - \sum_{i=1}^{n} \phi(\Delta_{(i)}, R_i(\beta_0), \hat{\Lambda}_n^{(\beta_0)}(Z_{(i)}))\right]$$

where ϕ is as defined in (2.5). For the sake of notational compactness, in the remainder of the proof, we will write $\hat{\Lambda}_n^{(\beta_0)}(Z_{(i)})$ as $\tilde{\Lambda}(Z_{(i)})$, $\hat{\Lambda}_{n,0}^{(\beta_0)}(Z_{(i)})$ as $\tilde{\Lambda}_0(Z_{(i)})$, and $\phi(\Delta_{(i)}, R_i(\beta_0), t)$ as $\phi_i(t)$.

Furthermore $\partial/\partial t \; \phi(\Delta_{(i)}, R_i(\beta_0), t)$ will be written as $\phi_i'(t)$ and so on. The set of indices i on which $\tilde{\Lambda}(Z_{(i)}$ and $\tilde{\Lambda}_0(Z_{(i)})$ differ is denoted by J_n. Now, $C_n = -2\, T_n$ where

$$
\begin{aligned}
T_n \;&=\; \sum_{i=1}^{n} \phi_i(\tilde{\Lambda}(Z_{(i)})) - \sum_{i=1}^{n} \phi_i(\tilde{\Lambda}_0(Z_{(i)})) \\
&=\; \sum_{i \in J_n} \phi_i(\tilde{\Lambda}(Z_{(i)})) - \sum_{i \in J_n} \phi_i(\tilde{\Lambda}_0(Z_{(i)})) \\
&=\; \sum_{i \in J_n} \phi_i^{'}(\Lambda_0(z_0))\, [(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0)) - (\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))] \\
&\qquad\qquad + \sum_{i \in J_n} \frac{1}{2}\, \phi_i^{''}(\Lambda_0(z_0))\; \left[(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0))^2 - (\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))^2\right] + R_n \\
&\equiv\; T_{n,1} + T_{n,2} + R_n\,,
\end{aligned}
$$

by Taylor–expanding $\phi_i(t)$ around $\Lambda_0(z_0)$. Here,

$$
R_n = \sum_{i \in J_n} \frac{1}{6}\, \phi_i^{'''}\,(\tilde{\Lambda}(Z_{(i)})^{\star})\; \left(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0)\right)^3 - \sum_{i \in J_n} \frac{1}{6}\, \phi_i^{'''}\,(\tilde{\Lambda}_0(Z_{(i)})^{\star})\; \left(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0)\right)^3
$$

(where $\tilde{\Lambda}(Z_{(i)})^{\star}$ is some point between $\tilde{\Lambda}(Z_{(i)})$ and $\Lambda_0(z_0)$ and $\tilde{\Lambda}_0(Z_{(i)})^{\star}$ is some point between $\tilde{\Lambda}_0(Z_{(i)})$ and $\Lambda_0(z_0)$) and can be shown to converge to 0 in probability by using the facts that (a) $\sup_{i \in J_n} \mid \phi_i^{'''}(\tilde{\Lambda}(Z_{(i)})^{\star}) \mid$ and $\sup_{i \in J_n} \mid \phi_i^{'''}(\tilde{\Lambda}_0(Z_{(i)})^{\star}) \mid$ are $O_p(1)$, (b) $\sup_{z \in D_n} \mid \tilde{\Lambda}(z) - \Lambda_0(z_0) \mid$ and $\sup_{z \in D_n} \mid \tilde{\Lambda}(z) - \Lambda_0(z_0) \mid$ are $O_p(n^{-1/3})$ where D_n is the set on which $\tilde{\Lambda}$ and $\tilde{\Lambda}_0$ differ, and (c) the length of D_n is $O_p(n^{-1/3})$. Now consider $T_{n,2}$. Once again, by Taylor expansion, we have

$$
\begin{aligned}
T_{n,2} \;&=\; \sum_{i \in J_n} \frac{1}{2}\, \phi_i^{''}(\Lambda_0(z_0))\; \left[(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0))^2 - (\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))^2\right] \\
&=\; \sum_{i \in J_n} \frac{1}{2} \phi_i^{''}(\tilde{\Lambda}(Z_{(i)}))[\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0)]^2 - \sum_{i \in J_n} \frac{1}{2} \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)}))[\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0)]^2 \\
&\qquad\qquad + o_p(1)\,. \qquad (6.20)
\end{aligned}
$$

Now consider,

$$
T_{n,1} = \sum_{i \in J_n} \phi_i^{'}(\Lambda_0(z_0))(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0)) - \sum_{i \in J_n} \phi_i^{'}(\Lambda_0(z_0))\,(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0)) \equiv S_1 - S_2\,.
$$

Consider the term S_2. Note that for each $i \in J_n$, we can write:

$$
\phi_i^{'}(\Lambda_0(z_0)) = \phi_i^{'}(\tilde{\Lambda}_0(Z_{(i)})) + (\Lambda_0(z_0) - \tilde{\Lambda}_0(Z_{(i)}))\, \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)})) + \frac{1}{2}\, \phi_i^{'''}(\tilde{\Lambda}_0(Z_{(i)})^{\star\star})(\Lambda_0(z_0) - \tilde{\Lambda}_0(Z_{(i)}))^2
$$

where $\tilde{\Lambda}_0(Z_{(i)})^{\star\star}$ is a point between $\tilde{\Lambda}_0(Z_{(i)})$ and $\Lambda_0(z_0)$. We then have,

$$\begin{aligned}
S_2 &= \sum_{i \in J_n} \left[\phi_i^{'}(\tilde{\Lambda}_0(Z_{(i)})) + (\Lambda_0(z_0) - \tilde{\Lambda}_0(Z_{(i)}))\, \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)})) + \frac{1}{2}\, \phi_i^{'''}(\tilde{\Lambda}_0(Z_{(i)})^{\star\star})(\Lambda_0(z_0) - \tilde{\Lambda}_0(Z_{(i)}))^2 \right] \\
&= \qquad\qquad \times(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0)) \\
&= \sum_{i \in J_n} \left[\phi_i^{'}(\tilde{\Lambda}_0(Z_{(i)})) + (\Lambda_0(z_0) - \tilde{\Lambda}_0(Z_{(i)}))\, \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)})) \right] (\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0)) + o_p(1) \\
&= -\sum_{i \in J_n} \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)})) \left[\tilde{\Lambda}_0(Z_{(i)}) - \frac{\phi_i^{'}(\tilde{\Lambda}_0(Z_{(i)}))}{\phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)}))} - \Lambda_0(z_0) \right] (\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0)) + o_p(1)\,,
\end{aligned}$$

where the fact that the term involving $\phi_i^{'''}$ is $o_p(1)$ is deduced by arguments similar to those needed to show that R_n is $o_p(1)$. Now, let $B_1^0, B_2^0, \ldots, B_r^0$ denote the level blocks for $\tilde{\Lambda}_0(Z_{(i)})$ that constitute J_n, with level values $w_1^0, w_2^0, \ldots, w_r^0$ and suppose that $w_l^0 = \Lambda_0(z_0) \equiv \theta_0$. Then,

$$\begin{aligned}
S_2 + o_p(1) &= -\sum_{j=1}^{r} \sum_{i \in B_j} \left[\phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)})) \left(\tilde{\Lambda}_0(Z_{(i)})) - \frac{\phi_i^{'}(\tilde{\Lambda}_0(Z_{(i)}))}{\phi_i^{'''}(\tilde{\Lambda}_0(Z_{(i)}))} \right) - \Lambda_0(z_0)\, \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)})) \right] \\
&\qquad\qquad \times(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0)) \\
&= -\sum_{j=1}^{r} \sum_{i \in B_j} \left[\phi_i^{''}(w_j^0) \left(w_j^0 - \frac{\phi_i^{'}(w_j^0)}{\phi_i^{''}(w_j^0)} \right) - \Lambda_0(z_0)\, \phi^{''}(w_j^0) \right] (w_j^0 - \Lambda_0(z_0)) \\
&= -\sum_{j \neq l} (w_j^0 - \Lambda_0(z_0)) \left[\sum_{i \in B_j} (\phi_i^{''}(w_j^0)\, w_j^0 - \phi_i^{'}(w_j^0)) - \Lambda_0(z_0) \sum_{i \in B_j} \phi_i^{''}(w_j^0) \right] \\
&= -\sum_{j \neq l} (w_j^0 - \Lambda_0(z_0)) \left[\left(\sum_{i \in B_j} \phi_i^{''}(w_j^0) \right) \left[\frac{\sum_{i \in B_j} (\phi_i^{''}(w_j^0)\, w_j^0 - \phi_i^{'}(w_j^0))}{\sum_{i \in B_j} \phi_i^{''}(w_j^0)} - \Lambda_0(z_0) \right] \right] \\
&= -\sum_{j \neq l} \sum_{i \in B_j} \phi_i^{''}(w_j^0)\, (w_j^0 - \Lambda_0(z_0))^2\,,
\end{aligned}$$

where this last step follows from the following observation: If B' is a level block for $\tilde{\Lambda}_0$ contained in J_n with level value w^0, then

$$w^{(0)} = \frac{\sum_{k \in B'} (w^{(0)}\, \phi_k^{''}(w^{(0)}) - \phi_k^{'}(w^{(0)}))}{\sum_{k \in B'} \phi_k^{''}(w^{(0)})\,.}$$

provided $w^{(0)} \neq \theta_0$. This is a direct consequence of the representation (2.13). It follows that

$$\begin{aligned}
S_2 + o_p(1) &= -\sum_{j \neq l} \sum_{i \in B_j} \phi_i^{''}(w_j^0)\,(w_j^0 - \Lambda_0(z_0))^2 \\
&= -\sum_{j=1}^{r} \sum_{i \in B_j} \phi_i^{''}(w_j^0)\,(w_j^0 - \Lambda_0(z_0))^2 \\
&= -\sum_{j=1}^{r} \sum_{i \in B_j} \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)}))\,(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))^2 \\
&= -\sum_{i \in J_n} \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)}))\,(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))^2\,.
\end{aligned}$$

It is similarly established (using (2.11)) that

$$S_1 + o_p(1) = -\sum_{i \in J_n} \phi_i^{''}(\tilde{\Lambda}(Z_{(i)}))\,(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0))^2\,.$$

It follows that

$$T_{n,1} = -\sum_{i \in J_n} \phi_i^{''}(\tilde{\Lambda}(Z_{(i)}))\,(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0))^2 + \sum_{i \in J_n} \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)}))\,(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))^2 + o_p(1)\,.$$

Now, on using (6.20) and the fact that R_n is $o_p(1)$ we get

$$\begin{aligned}
T_n &= T_{n,1} + T_{n,2} + o_p(1) \\
&= -\frac{1}{2} \sum_{i \in J_n} \phi_i^{''}(\tilde{\Lambda}(Z_{(i)}))\,(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0))^2 + \frac{1}{2} \sum_{i \in J_n} \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)}))\,(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))^2 + o_p(1)\,,
\end{aligned}$$

whence

$$\begin{aligned}
C_n &= -2\,T_n = \sum_{i \in J_n} \phi_i^{''}(\tilde{\Lambda}(Z_{(i)}))\,(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0))^2 - \sum_{i \in J_n} \phi_i^{''}(\tilde{\Lambda}_0(Z_{(i)}))\,(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))^2 + o_p(1) \\
&= \sum_{i \in J_n} \phi_i^{''}(\Lambda_0(Z_{(i)}))\,(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0))^2 - \sum_{i \in J_n} \phi_i^{''}(\Lambda_0(Z_{(i)}))\,(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))^2 + o_p(1)\,.
\end{aligned}$$

Now,

$$\phi_i^{''}(\Lambda_0(Z_{(i)})) = \frac{\Delta_{(i)} \exp\left[-e^{\beta_0^T X_{(i)}}\,\Lambda_0(Z_{(i)})\right]\, e^{2\,\beta_0^T X_{(i)}}}{\left(1 - \exp\left[-e^{\beta_0^T X_{(i)}}\,\Lambda_0(Z_{(i)})\right]\right)^2}\,,$$

whence

$$\begin{aligned}
C_n &= \sum_{i \in J_n} \frac{\Delta_{(i)} \exp\left[-e^{\beta_0^T X_{(i)}}\,\Lambda_0(Z_{(i)})\right]\, e^{2\,\beta_0^T X_{(i)}}}{\left(1 - \exp\left[-e^{\beta_0^T X_{(i)}}\,\Lambda_0(Z_{(i)})\right]\right)^2} \left[(\tilde{\Lambda}(Z_{(i)}) - \Lambda_0(z_0))^2 - (\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))^2\right] + o_p(1) \\
&= n^{1/3}\,(\mathbb{P}_n - P)\,\Psi_n(\delta, z, x) + n^{1/3}\,P\,\Psi_n(\delta, z, x) + o_p(1)
\end{aligned}$$

where $\mathbb{P}_n$ is the empirical measure of the observations $\{\Delta_i, Z_i, X_i\}_{i=1}^n$, P denotes the true underlying distribution of (Δ, Z, X), Ψ_n is the random function given by

$$\Psi_n(\delta, z, x) = \frac{\delta \exp\left[-e^{\beta_0^T x}\,\Lambda(z)\right]\, e^{2\,\beta_0^T x}}{\left(1 - \exp\left[-e^{\beta_0^T x}\,\Lambda(z)\right]\right)^2}\,\left[(n^{1/3}(\tilde{\Lambda}(z) - \Lambda_0(z_0)))^2 - (n^{1/3}\,(\tilde{\Lambda}_0(z) - \Lambda_0(z_0)))^2\right]\,1(z \in D_n)\,.$$

We are using operator notation here for expectations; thus $\mathbb{P}_n\,g$ denotes the expectation of g under the measure $\mathbb{P}_n$ and $P\,g$ denotes the expectation of g under the measure P. The function g is allowed to be a random function. Now,

$$n^{1/3}\,(\mathbb{P}_n - P)\,\Psi_n(\delta, z, x) = n^{-1/6}\,\sqrt{n}\,(\mathbb{P}_n - P)\,\Psi_n(\delta, z, x)\,.$$

Using the facts that (i) D_n is eventually contained in a set of the form $[z_0 - M\,n^{-1/3}, z_0 + M\,n^{-1/3}]$ with arbirtrarily high preassigned probability (ii) the processes U_n and V_n are $O_p(1)$ on compacts and monotone increasing, along with standard preservation properties of Donsker classes of functions, it can be argued that with arbitrarily high preassigned probability, the function $\Psi_n(\delta, x, z)$ lies in a Donsker class, whence it follows that $\sqrt{n}\,(\mathbb{P}_n - P)\,\Psi_n(\delta, z, x)$ is $O_p(1)$; consequently $n^{1/3}\,(\mathbb{P}_n - P)\,\Psi_n(\delta, z, x)$ is $O_p(n^{-1/6})$ and hence $o_p(1)$.

To find the asymptotic distribution of C_n we can therefore concentrate on the asymptotic distribution of

$$n^{1/3}\,P\,\Psi_n(\delta, z, x) = n^{1/3}\,P\,\left[\frac{\Delta \exp\left[-e^{\beta_0^T X}\,\Lambda_0(Z)\right]\, e^{2\,\beta_0^T X}}{\left(1 - \exp\left[-e^{\beta_0^T X}\,\Lambda_0(Z)\right]\right)^2}\,K_n(Z)\right]$$

where

$$K_n(Z) = \left[(n^{1/3}(\tilde{\Lambda}(Z) - \Lambda_0(z_0)))^2 - (n^{1/3}(\tilde{\Lambda}_0(Z) - \Lambda_0(z_0)))^2\right]\,1(Z \in D_n)\,.$$

We can then write,

$$n^{1/3}\,P\,\Psi_n(\delta, z, x) = n^{1/3}\,P\,\left[K_n(Z)\,E_{Z,X}\,\left[\frac{\Delta \exp\left[-e^{\beta_0^T X}\,\Lambda_0(Z)\right]\, e^{2\,\beta_0^T X}}{\left(1 - \exp\left[-e^{\beta_0^T X}\,\Lambda_0(Z)\right]\right)^2}\right]\right]\,.$$

Using the fact that

$$E(\Delta \mid Z, X) = 1 - \exp\left[-e^{\beta_0^T X}\,\Lambda_0(Z)\right]$$

we have

$$n^{1/3}\,P\,\Psi_n(\delta, z, x) = n^{1/3}\,P\,\left[K_n(Z)\,\left[\frac{\exp\left[-e^{\beta_0^T X}\,\Lambda_0(Z)\right]\, e^{2\,\beta_0^T X}}{\left(1 - \exp\left[-e^{\beta_0^T X}\,\Lambda_0(Z)\right]\right)}\right]\right] \equiv n^{1/3}\,P\,\left[K_n(Z)\,\xi(Z, X)\right]\,,$$

where

$$\xi(Z,X) = \frac{\exp\left[-e^{\beta_0^T X}\,\Lambda_0(Z)\right]\, e^{2\,\beta_0^T X}}{\left(1 - \exp\left[-e^{\beta_0^T X}\,\Lambda_0(Z)\right]\right)}\,.$$

Thus,

$$\begin{aligned} n^{1/3}\, P\,\Psi_n(\delta,z,x) &= n^{1/3}\, P\,\left[K_n(Z)\,\xi(Z,X)\right] \\ &= n^{1/3} \int_{D_n} K_n(z)\, E(\xi(Z,X)\ \mid Z = z)\, f_Z(z)\, dz \\ &= n^{1/3} \int_{\tilde{D}_n} K_n(z_0 + h\, n^{-1/3})\, w(z_0 + h\, n^{-1/3})\, f_Z(z_0 + h\, n^{-1/3})\, dh \end{aligned}$$

where $h = n^{1/3}\,(z - z_0)$, $\tilde{D}_n = n^{1/3}\,(D_n - z_0)$ and $w(z) = E(\xi(Z,X)\ \mid Z = z)$. Now note that,

$$K_n(z_0 + h\, n^{-1/3}) = (U_n^2(h) - V_n^2(h))\, 1\,(h \in \tilde{D}_n)$$

where $\tilde{D}_n$ is the set on which U_n and V_n differ. Now, note that w is continuous in z and is given by:

$$w(z) = \int \frac{\exp\left[-e^{\beta_0^T x}\,\Lambda_0(z)\right]\, e^{2\,\beta_0^T x}}{\left(1 - \exp\left[-e^{\beta_0^T x}\,\Lambda_0(z)\right]\right)}\, \frac{f(z,x)}{f_Z(z)}\, d\,\mu(x)\,.$$

On using the facts that $\tilde{D}_n$ is eventually contained with arbitrarily high probability in a compact set and the boundedness in probability of the processes U_n and V_n on compacts along with the continuity of the functions w and f_Z, we get,

$$n^{1/3}\, P\,\Psi_n(\delta,z,x) = \int w(z_0)\, f_Z(z_0)\, (U_n^2(h) - V_n^2(h))\, dh + o_p(1)\,.$$

But $C(z_0) = w(z_0)\, f_Z(z_0) = 1/a^2$ where a is as defined in Theorem 3.2. An application of Theorem 3.2 and Slutsky's theorem yields

$$n^{1/3}\, P\,\Psi_n(\delta,z,x) \to_d \frac{1}{a^2} \int \left((g_{a,b}(h))^2 - (g^0_{a,b}(z))^2\right)\, dh\,,$$

and the fact that

$$\frac{1}{a^2} \int \left((g_{a,b}(h))^2 - (g^0_{a,b}(z))^2\right)\, dh \equiv_d \int \left((g_{1,1}(h))^2 - (g^0_{1,1}(z))^2\right)\, dh \equiv \mathbb{D}$$

follows as a direct application of Lemma 3.1 followed by the change of variable theorem from calculus.

It remains to show that

$$\tilde{R}_n \equiv 2\,(l_n(\hat{\beta}_{0n}, \hat{\Lambda}_n) - l_n(\beta_0, \hat{\Lambda}_n^{\beta_0})) - 2(l_n(\beta_{n,0}, \hat{\Lambda}_{n,0}) - l_n(\beta_0, \hat{\Lambda}_{n,0}^{\beta_0}))$$

is $o_p(1)$. This is precisely $\mathrm{lrtbeta}_n - \mathrm{lrtbeta}_n^0$. From Theorem 3.1 we get:

$$\begin{aligned}
\mathrm{lrtbeta}_n - \mathrm{lrtbeta}_n^0 &= n\,(\hat{\beta}_n - \beta_0)^T\,\tilde{I}_0\,(\hat{\beta}_n - \beta_0) - n\,(\tilde{\beta}_n - \beta_0)^T\,\tilde{I}_0\,(\tilde{\beta}_n - \beta_0) + o_p(1) \\
&= n\,(\hat{\beta}_n - \tilde{\beta}_n)^T\,\tilde{I}_0\,(\hat{\beta}_n - \tilde{\beta}_n) + 2\,n\,(\tilde{\beta}_n - \beta_0)^T\,\tilde{I}_0\,(\hat{\beta}_n - \tilde{\beta}_n) + o_p(1) \\
&= \sqrt{n}\,(\hat{\beta}_n - \tilde{\beta}_n)^T\,\tilde{I}_0\,\sqrt{n}\,(\hat{\beta}_n - \tilde{\beta}_n) + 2\,\sqrt{n}\,(\tilde{\beta}_n - \beta_0)^T\,\tilde{I}_0\,\sqrt{n}(\hat{\beta}_n - \tilde{\beta}_n) + o_p(1) \\
&\equiv I_n + II_n + o_p(1)\,.
\end{aligned}$$

The fact that I_n is $o_p(1)$ follows from the observation that $\sqrt{n}\,(\hat{\beta}_n - \tilde{\beta}_n) = r_n - s_n$, which is $o_p(1)$ (by Theorem 3.1). The fact that II_n is $o_p(1)$ follows on using the facts that $\sqrt{n}\,(\hat{\beta}_n - \tilde{\beta}_n)$ is $o_p(1)$ and that $\sqrt{n}\,(\tilde{\beta}_n - \beta_0)$ is $O_p(1)$. □

References

Banerjee, M. (2000). *Likelihood Ratio Inference in Regular and Nonregular Problems.* Ph.D. dissertation, University of Washington.

Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *Ann. Statist.* **29**, 1699–1731.

Banerjee, M. and McKeague, I. W. (2003). Confidence sets for split points in decision trees. *Submitted paper*, available at www.stat.lsa.umich.edu/∼moulib/splitdec20.pdf

Banerjee, M. (2004). Likelihood based inference for monotone functions: Towards a general theory. URL: http://www.stat.lsa.umich.edu/∼moulib/wilks3.pdf

Bloch, D.A. and Silverman, B.W. (1997) Monotone discriminant functions and their applications in rheumatology. *J. Amer. Statist. Assoc.* **92**, 144–153.

Dykstra, R.L. and Robertson, T. (1982) An algorithm for isotonic regression for two or more independent variables. *Ann. Statist.* **10**, 708–711.

Etzioni, R, Pepe M, Longton, G, et.al. (1999) Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Med Decis Making, 1999*, 242-251.

Dunson, D.B., (2004) Bayesian isotonic regression for discrete outcomes. *Working Paper*, available at *http://ftp.isds.duke.edu/ WorkingPapers/03-16.pdf*

Friedman, J. and Tibshirani, R.. (1984) The monotone smoothing of scatterplots. *Technometrics.* **26**, 243–250.

Freitag, S. (2004) Confidence intervals for quantiles in monotone binary regression models with application to current status data . *In preparation.*

Ghosal, S., Sen, A. and Van der Vaart, A.W. (2000) Testing monotonicity of regression. *Ann. Statist.* **28**, 1054-1082.

Gijbels, I. and Heckman, N. (2000) Nonparametric testing for a monotone hazard function via normalized spacings. *Technical Report*, **195**, Statistics Department, Univ. of British Columbia.

Groeneboom, P. and Wellner J.A. (1992). *Information Bounds and Nonparametric Likelihood Estimation.* Birkhäuser, Basel.

Hall, P. and Heckman, N. (2000) Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.* **28**, 20–39.

Hall, P. and Huang, L.S. (2001) Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.* **29**, 624-647.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* **??**, Monographs on Statistics and Applied Probability. Chapman and Hall, London.

He, Xuming and Shi, P. (1998) Monotone B–spline smoothing. *J. Amer. Statist. Assoc.* **93**, 643–650.

Huang, Y. and Zhang, C. (1994). Estimating a monotone density from censored observations. *Ann. Statist.* **24**, 1256 – 1274.

Huang, J. (1994). *Estimation in Regression Models with Interval Censoring.* Ph.D. dissertation, University of Washington.

Huang, J. (1996). Efficient estimation for the Proportional Hazards Model with Interval Censoring. *Ann. Statist.* **24**, 540 – 568.

Huang, J. (2002). A note on estimating a partly linear model under monotonicity constraints. *Journal of Statistical Planning and Inference.* **107**, 343–351.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *J. Comput. Graph. Statist.* **7**, 310-321.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18**, 191-219.

Magnac, T. and Maurin, E. (2002) Available at http://www.crest.fr/seminaires/lmi/magnacmaurin03.pdf

Mammen, E. (1991) Estimating a smooth monotone regression function. *Ann. Statist* **19**, 724-740.

Mammen, E., Marron, J.S, Turlach, B.A. and Wand, M.P. (2001) A general projection framework for constrained smoothing. *Statist. Sci.*, **16**, 232-248.

Manski, C.F. and Tamer, E. (2002) Inference in regressions with interval data on a regressor or outcome *Econometrica*, **70**, 519–546.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, **37**, Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Murphy, S.A. and Van der Vaart, A.W. (1997). Semiparametric Likelihood Ratio Inference. *Ann. Statist.* **25**, 1471 – 1509.

Murphy, S. and Van der Vaart, A. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** , 449 - 465.

Ramsay, J.O. (1998). Estimating smooth monotone functions. *JRSS*-B **60** , 365-375.

Ramsay, J.O. and Silverman, B. (1997) *Functional Data Analysis* Springer

Robertson,T., Wright, F.T. and Dykstra, R.L. (1988). *Order Restricted Statistical Inference.* Wiley, New York

Shiboski, S. (1998). Generalized Additive Models for Current Status Data. *Lifetime Data Analysis.* **4**, 29 - 50.

Van der Geer, S. A. (1990) Estimating a regression function. *Ann. Statist.* **18**, 907–924.

Van der Vaart, A. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

Villalobos, M. and Wahba, G. (1987) Inequality–constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *J. Amer. Statist. Assoc* **82**, 239-248.

Wellner, J. (2001). Gaussian white noise models: some results for monotone functions. *Crossing Boundaries: Statistical Essays in Honor of Jack Hall*, IMS Lecture Notes-Monograph Series, Vol **43** (2003), 87 – 104. J.E. Kolassa and D. Oakes, editors.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika.* **89**, 39–48.