doi: 10.1111/j.1467-9469.2006.00499.x

A Semiparametric Binary Regression Model Involving Monotonicity Constraints

MOULINATH BANERJEE

Department of Statistics, University of Michigan

PINAKI BISWAS and DEBASHIS GHOSH

Department of Biostatistics, University of Michigan

ABSTRACT. We study a binary regression model using the complementary log–log link, where the response variable Δ is the indicator of an event of interest (for example, the incidence of cancer, or the detection of a tumour) and the set of covariates can be partitioned as (X, Z) where Z (real valued) is the primary covariate and X (vector valued) denotes a set of control variables. The conditional probability of the event of interest is assumed to be monotonic in Z, for every fixed X. A finite-dimensional (regression) parameter β describes the effect of X. We show that the baseline conditional probability function (corresponding to $X = 0$) can be estimated by isotonic regression procedures and develop an asymptotically pivotal likelihood-ratio-based method for constructing (asymptotic) confidence sets for the regression function. We also show how likelihood-ratio-based confidence intervals for the regression parameter can be constructed using the chi-square distribution. An interesting connection to the Cox proportional hazards model under current status censoring emerges. We present simulation results to illustrate the theory and apply our results to a data set involving lung tumour incidence in mice.

Key words: binary regression, Brownian motion, chi-square distribution, Cox model, current status data, greatest convex minorant, likelihood ratio statistic, non-regular problem

1. Introduction

The study of shape-restricted functions arises extensively in statistical modelling. 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 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.

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 & Silverman, 1997). It seems reasonable that during this period of time, children's heights are increasing with age. Monotonicity also plays an important role in econometrics. In modelling labour participation as a function of wage, for example, classical theory of supply and demand implies that increased wages should be associated with increased labour 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 non-parametric 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 non-parametric isotonic regression procedures. A rate of convergence for estimators of monotone regression

functions was established by van de Geer (1990). Mammen (1991) analysed 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 & Huang (2001). Hypothesis testing procedures involving monotonicity have been proposed by Ghosal *et al.* (2000), Gijbels & Heckman (2000) and Hall & Heckman (2000). Procedures focusing more on the algorithmic aspects of non-parametric regression with monotonicity constraints have been given by many authors, among them Dykstra & Robertson (1982), He & Shi (1998), Ramsay (1998) and Mammen *et al.* (2001).

The preceding literature deals with the problem of studying non-parametric 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 & 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 non-continuous outcomes have not been well-addressed in the literature.

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

$$h(\mu(X,Z)) = \beta^{\mathrm{T}} X, \tag{1}$$

where $\mu(X,Z) = P(\Delta = 1 \mid X, Z)$ and h can be taken to be a smooth monotone increasing function from $(0,1)$ to $(-\infty, \infty)$ and is called the 'link function'. 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. It is assumed that higher values of Z are associated with higher chances of an outcome ($\Delta = 1$). Here and in the sequel, we assume monotone to mean increasing (this entails no loss of generality, since if the dependence of the response on the covariate is monotone decreasing, we can change the sign of the covariate and work with the transformed covariate). To incorporate the effect of Z in the model, and to ensure the monotonicity of the conditional probability of an outcome in Z, we extend (1) in the following way:

$$h(\mu(X,Z)) = \beta^{\mathrm{T}} X + \rho(Z), \tag{2}$$

where $\rho(z)$ is some monotone function of z. Thus the non-parametric 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.

In this article, we consider the use of likelihood-ratio-inference-based methods in a semiparametric binary regression model of the type described in (2). For the sake of concreteness we focus on a particular link function – the cloglog link. Thus, (2) reduces to:

$$\log(-\log(1-\mu(X,Z))) = \beta^{\mathrm{T}} X + \rho(Z),$$

where ρ 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 maximum likelihood esimators (MLEs) 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 (PH) model with current status data. This is a problem that has been fairly well studied in the recent past; see e.g. Huang (1994, 1996), Murphy & van der Vaart (1997). Hence, many of the techniques and results from this model can be fairly easily adapted to our setting. Writing $\mu(X, Z) = g(\beta, X, Z)$ and $\Lambda(\beta, X, Z) = -\log(1 - g(\beta, X, Z))$ and $\Lambda(Z) = \exp(\rho(Z))$, the above model can easily be written as:

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

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

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

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, $g(\beta, X, \cdot)$ can be thought of as a distribution function on the positive half-line. We call $g(\beta, 0, \cdot) \equiv g(\cdot)$, the baseline conditional probability function. The function $\Lambda(Z)$ is the cumulative hazard function corresponding to $g(Z)$ whereas $\Lambda(\beta, X, Z)$ is the cumulative hazard function corresponding to $g(\beta, X, Z)$.

1.1. The likelihood for a single observation and connections to the proportional hazards model

The density function of the vector (Δ, X, Z) can be written as:

$$p_{\beta,\Lambda}(\delta, x, z) = (1 - \exp(-\Lambda(z)\exp(\beta^{\mathrm{T}} x)))^{\delta}\,(\exp(-\Lambda(z)\exp(\beta^{\mathrm{T}} x)))^{1-\delta} f(x, z), \tag{3}$$

where $f(x, z)$ is the joint density of (X, Z) 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 (3) is identical to that for the Cox PH 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, W, Y), where $D = 1(T \leq Y)$.

Let $\Lambda(t \mid w)$ denote the conditional cumulative hazard function of the survival time T given $W = w$. Suppose that

$$\Lambda(t \mid w) = \Lambda(t)\exp(\beta^{\mathrm{T}} w).$$

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

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

Substituting

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

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

$$\tilde{p}(d, w, y) = (1 - \exp(-\Lambda(y)\exp(\beta^{\mathrm{T}} w)))^{\delta} (\exp(-\Lambda(y)\exp(\beta^{\mathrm{T}} w)))^{1-\delta} f(w, y). \tag{4}$$

Comparing (4) with (3), we find that they are identical. This implies that the joint distribution of (D, W, Y) is the same as the joint distribution of (Δ, X, Z). Thus the sample $\{\Delta_i, X_i, Z_i\}_{i=1}^n$ 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 ith 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 MLEs, 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 regions 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 MLEs would yield) is well known; see the discussion in the introduction of Murphy & 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 existing results, 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 non-regular statistical problems involving $n^{1/3}$ rate of convergence for MLEs and non-Gaussian limit distributions. Indeed the MLE $\hat{\Lambda}_n$ 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 nuisance parameters. 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). From the point of view of applications, it is also of interest to be able to estimate the regression function (or the conditional probability function) in this model. We will show in section 3 that confidence sets for the regression function $\mu(x, z)$ can also be constructed using likelihood-ratio-based inversion (as in the case of the baseline probability function), with calibration given by the quantiles of $\mathbb{D}$ (as in the baseline scenario).

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

mice tumour study in section 4. We conclude with some discussion in section 5. Proofs of some of the results in section 3 are collected in the appendix (section 6).

2. Computing MLEs 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,

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

Let $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}$ denote the ordered values of the Z_is; let $\Delta_{(i)}$ and $X_{(i)}$ denote the indicator and covariate values associated with value $Z_{(i)}$. Also, let $\Lambda_i \equiv \Lambda(Z_{(i)})$ and $R_i(\beta) = \exp(\beta^{\mathrm{T}} X_{(i)})$. For $\delta \in \{0, 1\}$ and $r, u \geq 0$ set,

$$\phi(\delta, r, u) = -\delta \log(1 - \mathrm{e}^{-ru}) + (1 - \delta) ru. \tag{5}$$

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

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

Minimizing Ψ with respect to β and Λ amounts to finding

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

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 $\{Z_{(i)}\}_{i=1}^{n}$. For $z < Z_{(1)}$, we set $\hat{\Lambda}_n(z) = 0$.

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

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

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

$$\mathrm{lrtbeta}_n = 2(l_n(\hat{\beta}_n, \hat{\Lambda}_n) - l_n(\beta_0, \hat{\Lambda}_n^{(\beta_0)})). \tag{6}$$

We next discuss the computation of the constrained maximizers of $l_n(\beta, \Lambda)$, 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

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

Then,

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

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

$$\mathrm{lrtLambda}_n = 2(l_n(\hat{\beta}_n, \hat{\Lambda}_n) - l_n(\hat{\beta}_{n,0}, \hat{\Lambda}_{n,0})). \tag{7}$$

One way of computing the MLEs 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 MLEs 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

$$\frac{\partial}{\partial\beta}\Psi(\beta,\Lambda) = 0. \tag{8}$$

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 (8). However, one can try to find a zero of (8) in the set $\{(\beta, \hat{\Lambda}_n^{(\beta)}) : \beta \text{ varies}\}$. 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 satisfy (8) in general. However, $\hat{\beta}_n(\hat{\Lambda}_n) = \hat{\beta}_n$, so we are guaranteed at least one solution, namely the MLEs of β and Λ. Though we were not able to establish that any root of (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

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

in the following manner:

(1) Choose an initial value $\beta^{(0)}$ and a small number ϵ.
(2) Set $\beta = \beta^{(0)}$. Compute $\hat{\Lambda}_n^{(\beta)}$.
(3) Solve,

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

Set $\beta^{(0)}$ to be equal to the solution. If $|\beta^{(0)} - \beta| < \epsilon$, stop. Otherwise go to Step (2).

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

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

2.1. *Characterizing* $\hat{\Lambda}_n^{(\beta)}$

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

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

over all $0 \le u_1 \le u_2 \le \cdots \le u_n$. 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 minimizing 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:

$$\sum_{j=i}^{n} \frac{\partial \phi(\Delta_{(j)}, R_j(\beta), u_j)}{\partial u_j} \left(\hat{u}_j\right) \ge 0 \quad \text{for} \quad i = 1, 2, \ldots, n \tag{9}$$

and

$$\sum_{j=1}^{n} \hat{u_j} \frac{\partial \phi(\Delta_{(j)}, R_j(\beta), u_j)}{\partial u_j} \left(\hat{u}_j\right) = 0. \tag{10}$$

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

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

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 (2005). We omit the details here but provide the 'self-induced' characterization of $\hat{u}$. For $1 \le i \le n$, 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 $0 \le u_1 \le u_2 \le \cdots \le u_n$ and hence furnishes the isotonic regression of the function

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

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

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

where slogcm$[\{x_i, y_i\}_{i=0}^n]$ (with $x_0 = y_0 = 0$ and $x_0 < x_1 < \cdots < x_n$) denotes the vector of slopes (left-derivatives) of the greatest convex minorant (GCM) of the piecewise linear function L [obtained by joining successive points $(x_i, y_i), (x_{i+1}, y_{i+1})$ by straight lines], computed at the points $x_1, x_2, \ldots, x_n$. 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 $u^{(0)}$, 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 $r(i)$ in the current situation. We have

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

and

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

with

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

The algorithm stops when the Fenchel conditions (9) and (10) are satisfied to a pre-specified degree of tolerance.

An important consequence of the above self-induced 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 $r(j)$s for the js in that block, with the weights given by the d_js. 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

$$\hat{\Lambda}_n^{(\beta)}(Z_{(j)}) = \frac{\sum_{k \in B_i} \{\hat{\Lambda}_n^{(\beta)}(Z_{(k)}) \, \phi''_{k,\beta}(\hat{\Lambda}_n^{(\beta)}(Z_{(k)})) - \phi'_{k,\beta}(\hat{\Lambda}_n^{(\beta)}(Z_{(k)}))\}}{\sum_{k \in B_i} \phi''_{k,\beta}(\hat{\Lambda}_n^{(\beta)}(Z_{(k)}))} \quad \text{for} \quad j \in B_i. \tag{11}$$

2.2. *Characterizing* $\hat{\Lambda}_{n,0}^{(\beta)}$

Let m be the number of Z_is 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 $0 \le u_1 \le u_2 \cdots \le u_m \le \theta_0 \le u_{m+1} \le \cdots \le u_n$. This can be reduced to solving two separate optimization problems. These are:

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

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

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 $0 \le u_1 \le u_2 \le \cdots \le u_m$ 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

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

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

$$\hat{\Lambda}_n^{(\beta)}(z) \ne \hat{\Lambda}_{n,0}^{(\beta)}(z) \Rightarrow \hat{\Lambda}_{n,0}^{(\beta)}(z_0) = \theta_0 \quad \text{or} \quad \hat{\Lambda}_n^{(\beta)}(z) = \hat{\Lambda}_n^{(\beta)}(z_0). \tag{12}$$

The constrained solution also has a 'self–induced' characterization in terms of the slope of the GCM 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-induced characterization.

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

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

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

$$r_0(i) = \hat{u}_i^{(0)} - \bigtriangledown_i \psi(\hat{u}^{(0)})\, d_{i,0}^{-1}$$

on the ordered set $\{1, 2, \ldots, n\}$ with weight function $d_{i,0} \equiv \bigtriangledown_{ii} \psi(\hat{u}^{(0)})$. 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-induced' 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 $\Sigma_{i \in \tilde{B}} r_0(i)\, d_{i,0} / \sum_{i \in \tilde{B}} d_{i,0}$. 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 $\tilde{w}_i \ne \theta_0$, it can be written as

$$\tilde{w}_i \equiv \hat{\Lambda}_{n,0}^{(\beta)}(Z_{(j)}) = \frac{\sum_{k \in \tilde{B}_i} \{ \hat{\Lambda}_{n,0}^{(\beta)}(Z_{(k)})\, \phi''_{k,\beta}(\hat{\Lambda}_{n,0}^{(\beta)}(Z_{(k)})) - \phi'_{k,\beta}(\hat{\Lambda}_{n,0}^{(\beta)}(Z_{(k)})) \}}{\sum_{k \in \tilde{B}_i} \phi''_{k,\beta}(\hat{\Lambda}_{n,0}^{(\beta)}(Z_{(k)}))} \quad \text{for} \quad j \in \tilde{B}_i. \tag{13}$$

This representation will prove useful later on.

3. Asymptotic results

In this section we present asymptotic results for the estimation of β and Λ. The parameter space for β is taken to be a bounded subset of $\mathbb{R}^d$. We denote it by $\mathcal{C}$. The parameter space for Λ is the space of all non-decreasing 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 $P(\|X\| \leq x_0) = 1$. Also $E(\mathrm{Var}(X \mid Z))$ is positive definite with probability 1.
(A.3) Let g_0 denote the true baseline conditional probability function (so, $\Lambda_0 = -\log(1 - g_0)$). 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 MLEs (see p. 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 positive 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 Z. The assumption (A.6) is a very technical assumption and is required to ensure that one can define appropriate approximately *least favourable submodels* as in Murphy & van der Vaart (1997; pp. 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^{**} given by (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, X, Z) is given by:

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

The ordinary score function for β in this model is:

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

where

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

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:

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

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

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

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

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

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

$$h^*(Z) = \Lambda_0(Z)h^{**}(Z) = \Lambda_0(Z)\frac{E_{\beta_0,\Lambda_0}(ZQ^2((\Delta, X, Z); \beta_0, \Lambda_0 \mid Z))}{E_{\beta_0,\Lambda_0}(Q^2((\Delta, X, Z); \beta_0, \Lambda_0 \mid Z))}. \tag{14}$$

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 $\tilde{I}_0 = \mathrm{Disp}(\tilde{l}) \equiv E_{\beta_0,\Lambda_0}(\tilde{l}\tilde{l}^{\mathrm{T}})$ is positive definite (note that $E_{\beta_0,\Lambda_0}(\tilde{l}) = 0$). This ensures 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:

$$\mathrm{lrtbeta}_n^0 = 2 \log \frac{\mathrm{argmax}_{\Lambda(z_0)=\theta_0}\, l_n(\beta, \Lambda)}{\mathrm{argmax}_{\beta=\beta_0,\Lambda(z_0)=\theta_0}\, l_n(\beta, \Lambda)}. \tag{15}$$

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 behaviour 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 (6) and $\mathrm{lrtbeta}_n^0$ above.

Theorem 1

Under conditions (A.1)–(A.6), both $\hat{\beta}_n$ and $\tilde{\beta}_n$ are asymptotically linear in the efficient score function and have the following representation:

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

and

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

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,

$$lrtbeta_n = n(\hat{\beta}_n - \beta_0)^{\mathrm{T}} \tilde{I}_0 (\hat{\beta}_n - \beta_0) + o_p(1), \tag{16}$$

while

$$lrtbeta_n^0 = n(\tilde{\beta}_n - \beta_0)^{\mathrm{T}} \tilde{I}_0 (\tilde{\beta}_n - \beta_0) + o_p(1). \tag{17}$$

It follows that both $lrtbeta_n$ and $lrtbeta_n^0$ are asymptotically 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 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 PH model with current status data in Murphy & van der Vaart (1997). The derivation in Murphy & van der Vaart (1997) is done for a one-dimensional β but the proof extends easily to higher dimensions. The asymptotically linear representation for $\tilde{\beta}_n$ and the limiting chi-square distribution for 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 lrtbeta_n, is asymptotically chi-square. 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 non-parametric component of the model. In order to do so, we introduce the following processes. For positive constants c and d define the process $X_{c,d}(z) := cW(z) + dz^2$, where $W(z)$ is standard two-sided Brownian motion starting from 0. Let $G_{c,d}(z)$ denote the GCM 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) neighbourhood of 0, whereas $g_{c,d}(h)$ is almost surely non-zero in some (random) neighbourhood 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 & Wellner (2001) and Wellner (2003). 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 1

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

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

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:

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

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

Theorem 2
Define,

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

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

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

where λ_0 *is the derivative of* Λ_0. *The processes* $(U_n(h), V_n(h))$ *converge finite dimensionally to the processes* $(g_{a,b}(h), g^0_{a,b}(h))$. *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 behaviour of the likelihood ratio statistic for testing (the true) $\tilde{H}_0 : \Lambda(z_0) = \theta_0$.

Theorem 3
The likelihood ratio statistic for testing $\tilde{H}_0 : \Lambda(z_0) = \theta_0$ *as defined in (7) converges in distribution to* $\mathbb{D}$ *where*

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

3.1. Construction of confidence sets for parameters of interest via likelihood-ratio-based inversion

Denote the likelihood ratio statistic for testing the null hypothesis $\Lambda(z_0) = \theta$ by $\mathrm{lrtLambda}_n(\theta)$. The computation of the likelihood ratio statistic is discussed, in detail, in section 2. By theorem 3, an approximate level $1 - \alpha$ confidence set for $\Lambda_0(z_0)$ is given by $S_{\Lambda_0(z_0)} \equiv \{\theta : \mathrm{lrtLambda}_n(\theta) \le q(\mathbb{D}, 1 - \alpha)\}$, where $q(\mathbb{D}, 1 - \alpha)$ is the $(1 - \alpha)$th quantile of the distribution of $\mathbb{D}$ (for $\alpha = 0.05$, this is approximately 2.28). Furthermore, the corresponding confidence set for the baseline conditional probability function g_0 at the point z_0, i.e. $P(\Delta = 1 \mid X = 0, Z = z_0)$ is simply $1 - \mathrm{e}^{-S_{\Lambda_0(z_0)}}$.

Confidence sets for the regression function at values $X = x_0, Z = z_0$, i.e. $\mu(x_0, z_0) = P(\Delta = 1 \mid X = x_0, Z = z_0)$ can also be constructed in a similar fashion. This requires simply re-defining the covariate X, so as to convert $\mu(x_0, z_0)$ to a baseline conditional probability. Set $\tilde{X} = X - x_0$. Then $\mu(x_0, z_0) = P(\Delta = 1 \mid \tilde{X} = 0, Z = z)$. Define $\tilde{\mu}(\tilde{x}, z) = E(\Delta = 1 \mid \tilde{X} = \tilde{x}, Z = z)$. We have,

$$\tilde{\mu}(\tilde{x}, z) = \mu(\tilde{x} + x_0, z) = 1 - \exp[-\Lambda_0(z)\, \mathrm{e}^{\beta_0^{\mathrm{T}}(\tilde{x} + x_0)}] = 1 - \exp[-\tilde{\Lambda}_0(z)\, \mathrm{e}^{\beta_0^{\mathrm{T}} \tilde{x}}],$$

where $\tilde{\Lambda}_0(z) = e^{\beta_0^{\mathrm{T}} x_0} \Lambda(z)$, with $\tilde{\rho}_0(Z) = \log \tilde{\Lambda}_0(Z)$. Note that $\tilde{\Lambda}(0) = 1$. But this is exactly the binary regression model considered in section 1 and satisfies the regularity conditions in section 3. Now, $\mu(x_0, z_0) = \tilde{\mu}(0, z_0) = 1 - \mathrm{e}^{-\tilde{\Lambda}_0(z_0)}$. An approximate level $1 - \alpha$ confidence set for $\tilde{\Lambda}_0(z_0)$, say $\tilde{S}_{\tilde{\Lambda}_0(z_0)}$ can be found in exactly the same fashion as before; i.e. $\tilde{S}_{\tilde{\Lambda}_0(z_0)} = \{\theta : \widetilde{\mathrm{lrtLambda}}_n(\theta) \le q(\mathbb{D}, 1 - \alpha)\}$, where $\widetilde{\mathrm{lrtLambda}}_n(\theta)$ is the likelihood ratio statistic for testing $\tilde{\Lambda}(z_0) = \theta$ and is computed in exactly the same way as the statistic in (7), but using the covariates $\tilde{X}$ and Z, instead of X and Z. Correspondingly, the confidence set for $\tilde{\mu}(0, z_0)$ is $1 - \mathrm{e}^{-\tilde{S}_{\tilde{\Lambda}_0(z_0)}}$. By the correspondence with the Cox PH model under current status censoring

(as discussed in section 1), this technique can be employed to construct an asymptotic level $1-\alpha$ confidence set for $F(y \mid w)$, the (conditional) distribution function of the survival time T at a fixed point y given that the covariate W assumes the value w. We apply this principle to construct confidence sets for the conditional probabilities in the data analysis example in section 4.

Confidence sets for the finite-dimensional regression parameter β_0 can be constructed in the usual fashion as: $\{\beta : \text{lrtbeta}_n(\beta) \leq q_{\chi^2_d, 1-\alpha}\}$, where $\text{lrtbeta}_n(\beta)$ is the likelihood ratio statistic for testing the null hypothesis that the true regression parameter is β (see (6)), and $q_{\chi^2_d,1-\alpha}$ is the $(1-\alpha)$th quantile of the χ^2_d distribution.

4. Simulation studies and data analysis

In an attempt to assess the performance of the techniques described above, extensive simulation studies were conducted of which only a part will be displayed here for the sake of brevity.

Data was generated as $\{(\Delta_i, X_i, Z_i) : i = 1, 2, \ldots, n\}$ from $p_{\beta,\Lambda}(\cdot)$, as defined in (3), 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 (X_i, Z_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, but truncated to lie in $[-2, 2]$, whereas Z_i had an exponential distribution with mean 1, truncated to [0.5, 1.5]. In the second case, X_i had the same distribution as in the first case but conditional on X_i, Z_i had a truncated exponential distribution on [0.5, 1.5] with mean $8/(8 + X_i)$. The conditional distribution of Δ_i was thus Bernoulli, with probability $p_i = 1 - \exp\{-\Lambda(Z_i)e^{\beta X_i}\}$. It is easy to verify that the above setup satisfies the assumptions (A.1)–(A.6) in section 3.

The sample sizes chosen were 200, 500, 1000, while β assumed the values -0.5, -0.25, 0.0, 0.25 and 0.5. The choice for $\Lambda(t)$ was taken as $\Lambda(t) = t^4$, which is convex. Simulations for a linear and concave $\Lambda(t)$ were also carried out but we shall omit them here on account of similar performance.

Tables 1 and 2 demonstrate the performance of the confidence intervals (CI) so obtained where for each fixed value of β, the first row refers to the coverage probability (target = 97.5%), with the second row showing the average length of the interval. These are computed, based on 1000 replicates, for each situation. As expected, the intervals get narrower with increasing sample size. The coverage for $\Lambda(z_0)$ is not affected by changes in the value of β although the coverage for β gets affected as one moves farther from 0.

Table 1. *Simulation results when Z_i and X_i are independent, with $z_0 = 1.0$ and $x_0 = 1.0$*

	$n = 200$		$n = 500$		$n = 1000$	
β	CI(β)	CI($\Lambda(z_0)$)	CI(β)	CI($\Lambda(z_0)$)	CI(β)	CI($\Lambda(z_0)$)
-0.5	94.7	97.7	96.3	96.9	96.0	96.8
	0.782	1.256	0.455	0.852	0.311	0.654
-0.25	96.4	95.7	96.3	96.4	96.7	96.7
	0.727	1.178	0.424	0.843	0.290	0.651
0.0	97.1	97.5	97.3	96.6	97.5	97.5
	0.703	1.199	0.413	0.842	0.284	0.641
0.25	95.2	96.7	97.0	98.4	96.9	96.7
	0.726	1.224	0.424	0.860	0.290	0.651
0.5	94.6	96.8	94.9	96.3	95.5	97.4
	0.776	1.244	0.455	0.871	0.312	0.655

Table 2. *Simulation results when Z_i and X_i are dependent, with $z_0 = 1.0$ and $x_0 = 1.0$*

	$n = 200$		$n = 500$		$n = 1000$	
β	CI(β)	CI($\Lambda(z_0)$)	CI(β)	CI($\Lambda(z_0)$)	CI(β)	CI($\Lambda(z_0)$)
−0.5	93.9	95.7	94.8	97.8	95.3	96.4
	0.775	1.272	0.456	0.871	0.312	0.662
−0.25	96.7	97.2	96.3	98.1	95.6	97.4
	0.721	1.214	0.426	0.842	0.291	0.644
0.0	97.0	97.3	96.8	96.7	97.2	97.0
	0.709	1.191	0.413	0.832	0.283	0.648
0.25	96.4	97.7	96.2	97.2	97.2	97.7
	0.729	1.206	0.425	0.845	0.291	0.651
0.5	94.9	96.9	97.6	97.1	96.1	97.0
	0.776	1.252	0.454	0.873	0.311	0.661

Figure 1 demonstrates the performance of the unconstrained and constrained non-parametric estimators of $\Lambda(t)$ for a simulated sample of size 1000 and $z_0 = 1.0$, in the situation that X and Z are independent and $\beta = 0.25$. The estimators are doing quite well, and differing only in a neighbourhood of $z_0 = 1.0$.

Figure 2 displays the quantile–quantile plot for the likelihood ratio statistic for testing $\tilde{H}_0$ at different sample sizes (in the underlying simulation setting $\beta = 0.25$, X and Z are independent and $z_0 = 1.0$) 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 centred and scaled MLE of β, in finite samples. This is carried out by regressing selected quantiles of the likelihood ratio statistic on the corresponding quantiles

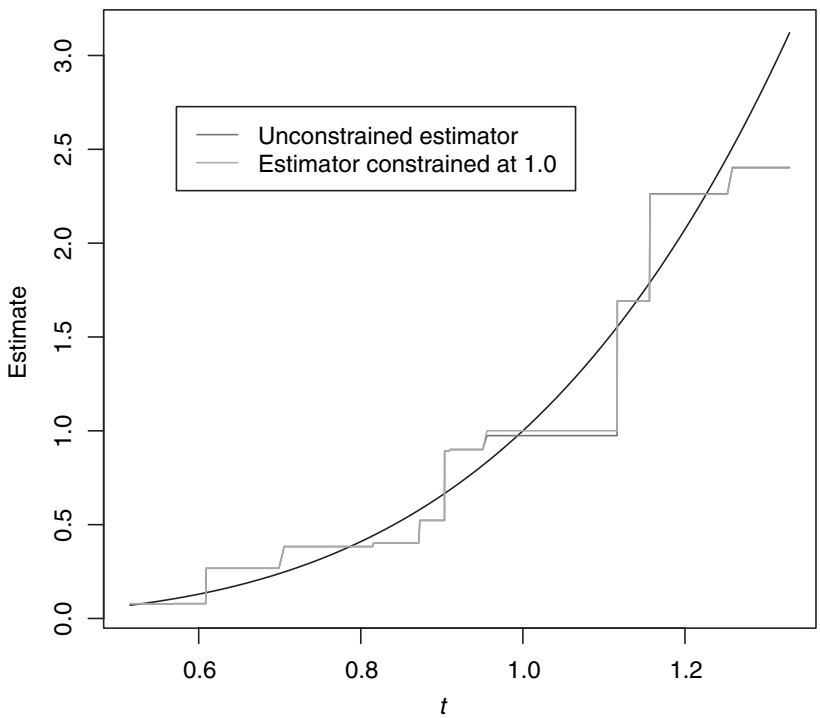

Fig. 1. Constrained and unconstrained non-parametric estimators of $\Lambda(t)$ obtained from simulated data.

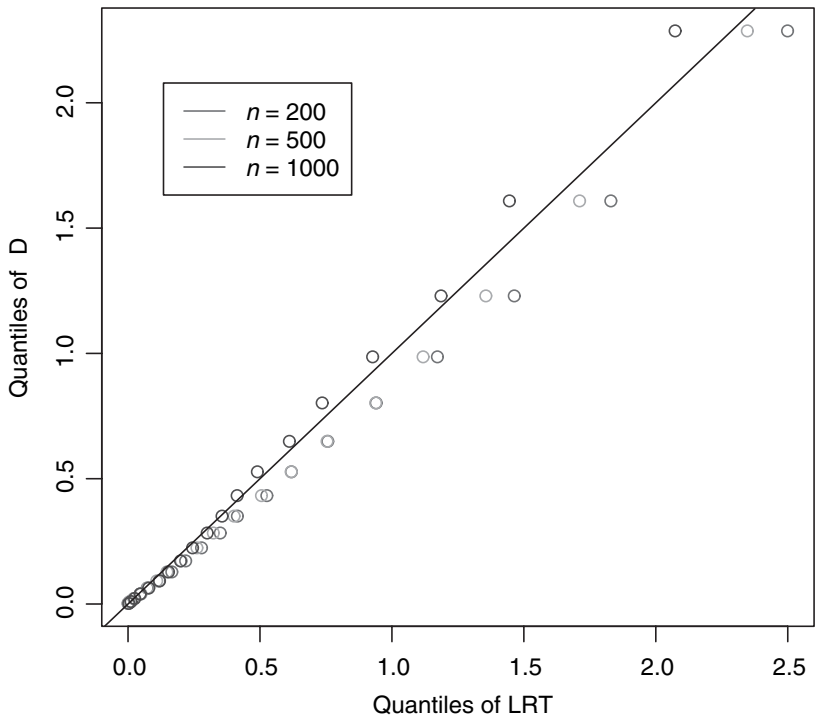

Fig. 2. Performance of the likelihood ratio statistic for testing $\tilde{H}_0$ when Z is independent of X.

Table 3. *Performance of the likelihood ratio test (LRT) statistic versus the centred ML statistic*

Effect	Estimate	SE	t-value	P-value
Linear regression of LRT(β) quantiles on χ^2_1 quantiles				
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				
Intercept	0.5940	0.0122	48.54	<2e-16
Slope	2.2243	0.0142	157.03	<2e-16

of the χ^2_1 distribution and by regressing the quantiles of $\sqrt{n}(\hat{\beta}-\beta)$ on those 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. As far as Table 3 is concerned, the underlying simulation setting had $\beta=0.25$, X and Z independent and $z_0=1$.

4.1. Data analysis

The methods of this paper were applied to a data set from Hoel & Walburg (1972), see also Finkelstein & Wolfe (1985) and Finkelstein (1986). A total of 144 RFM mice were randomly assigned to either a germ-free or conventional environment. The purpose of the study was to compare the time until lung tumour onset (for these two groups). Each mouse was inspected for lung tumour at the time of its death. However, the (random) time to tumour onset (T) is not directly observable, since lung tumour is non-lethal in mice. Thus, the only observables are the (random) duration between the start of the study and death (Y, measured in days),

and the indicator of whether a tumour is present at time of death. We are therefore in the set-up of a current status model, with T interval-censored by Y; we only observe $D = 1(T \leq Y)$, and the covariate W is the indicator of the conventional group. Note here, that if the tumour were rapidly lethal, it would be reasonable to treat the ages at death as failure times if the tumour is present, and to treat them as censoring times if not (interesting focusing in this case, on the time to death as a consequence of the lethal tumour). This would lead to right-censored data. We fit the Cox PH model to the data (as discussed in section 1); we denote by F_1 the distribution function of the time to tumour development in the conventional group and by F_0 the corresponding quantity for the time to tumour development in the germ-free group. This is of course tantamount to fitting the binary regression model in this paper with $\Delta \equiv D$, $Z \equiv Y$ and $X \equiv W$.

The estimate of the regression parameter (the effect of environment) was obtained as $\hat{\beta}_n = -0.654$. The confidence interval for β, based on the likelihood ratio test was obtained as $[-1.054, -0.204]$. This indicates that the conventional mice have a higher probability of 'surviving' longer than germ-free mice; in other words F_1 is dominated by F_0. This is illustrated in Fig. 3, where Λ_1, the cumulative hazard function corresponding to F_1 is plotted along with Λ_0, the cumulative hazard corresponding to F_0.

Table 4 shows pointwise confidence intervals for F_0 and F_1 at selected values of z [these are simply the regression functions for the two different groups of mice; $F_0(z) = P(\Delta = 1 \mid X = 0, Z = z)$ and $F_1(z) = P(\Delta = 1 \mid X = 1, Z = z)$] obtained using likelihood ratio tests, as described in section 3, and also the 'model-based bootstrap'. The model-based bootstrap was carried out by resampling Δ^* from the conditional distribution of Δ given the values of age at death and group (obtained from the given data). The estimation procedure was carried

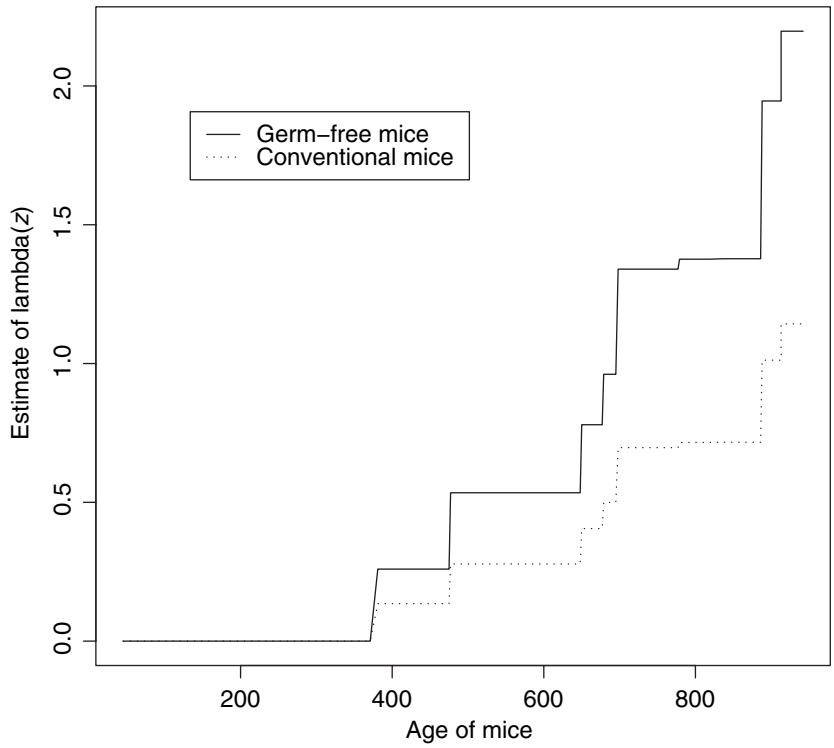

Fig. 3. Estimate of $\Lambda(z)$ for germ-free and conventional mice.

Table 4. *Confidence intervals across ages for the regression function in germ-free and conventional mice*

	Normal mice ($X=1$)			Germ-free mice ($X=0$)		
Age	$\hat{p}_n$	LRT	Bootstrap	$\hat{p}_n$	LRT	Bootstrap
450	0.126	0.039–0.288	0.000–0.252	0.228	0.039–0.589	0.000–0.457
500	0.243	0.131–0.323	0.000–0.485	0.414	0.173–0.646	0.000–0.828
600	0.243	0.173–0.356	0.108–0.377	0.414	0.213–0.646	0.150–0.678
700	0.502	0.213–0.664	0.289–0.715	0.738	0.323–0.870	0.341–1.000
800	0.511	0.356–0.763	0.238–0.784	0.747	0.523–0.888	0.539–0.956
900	0.637	0.387–0.925	0.273–1.000	0.857	0.646–0.969	0.714–1.000

out on the resampled data and new estimates of the regression function $\hat{p}_n^*$ were computed after which one obtained the percentiles from the bootstrap distribution of $n^{1/3}\,|\,\hat{p}_n^* - \hat{p}_n\,|$ (500 resamplings were used for this purpose). Note that the conventional bootstrap is highly suspect in problems with $n^{1/3}$ convergence rates, but the model-based bootstrap can be expected to work (there is however, no rigorous justification for it at this point). Based on some preliminary simulations, likelihood-ratio-based intervals are seen to be shorter than the ones based on the model-based bootstrap, but with comparable coverage. Also, the likelihood-ratio-based intervals are theoretically validated. We therefore advocate the use of the likelihood-ratio-based CIs for the regression function.

5. Discussion

In this paper, we have studied a semiparametric binary regression model with a natural connection to the Cox PH model under current status censoring and applied it to studying the time to development of tumour in mice. The effect of the auxiliary covariates is captured by a finite-dimensional regression parameter, whereas that of the principal covariate is specified through a monotone increasing function. While we have used the complementary log–log link in our modelling scheme, similar results should hold for link functions that preserve the concavity of the log-likelihood function in Λ and are adequately differentiable. 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. Furthermore, (asymptotic) likelihood-ratio-based confidence sets of any pre-assigned level of confidence for the regression function can also be constructed by a simple shift transformation on the auxiliary covariates X, as illustrated in section 3.

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, 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 is 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 analysed 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.

Acknowledgement

The first author's research was partly supported by a grant from the National Science Foundation.

References

Banerjee, M. (2000). *Likelihood ratio inference in regular and nonregular problems*. PhD Dissertation, University of Washington, Washington.

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

Banerjee, M. (2005). Likelihood based inference for monotone response models. URL: http://www.stat.lsa.umich.edu/∼moulib/wilks-rectify-2.pdf.

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

Finkelstein, D. M. (1986). A proportional hazards model for interval censored failure time data. *Biometrics*, **42**, 845–854.

Finkelstein, D. M. & Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval censored failure time data. *Biometrics* **41**, 933–945.

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

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

Groeneboom, P. (1989). Brownian motion with a parabolic drift and Airy functions. *Probab. Theory Relat. Fields* **81**, 79–109.

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

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

Hastie, T. & Tibshirani, R. (1990). *Generalized additive models*. Monographs on Statistics and Applied Probability. Chapman and Hall, London.

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

Hoel, D. G. & Walburg, H. E. (1972). Statistical analysis of survival experiments. *J. Natl. Cancer Inst.* **49**, 361–372.

Huang, Y. & 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*. PhD Dissertation, University of Washington, 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. *J. Statist. Plann. Inference* **107**, 343–351.

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

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

Mammen, E., Marron, J. S., Turlach, B. A. & Wand, M. P. (2001). A general projection framework for constrained smoothing. *Statist. Sci.* **16**, 232–248.
Murphy, S. A. & van der Vaart, A. W. (1997). Semiparametric likelihood ratio inference. *Ann. Statist.* **25**, 1471–1509.
Ramsay, J. O. (1998). Estimating smooth monotone functions. *J. Roy. Statist. Soc. Ser. A* **60**, 365–375.
Ramsay, J. O. & Silverman, B. (1997). *Functional data analysis*. Springer, New York.
Robertson, T., Wright, F. T. & Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley, New York.
Shiboski, S. (1998). Generalized additive models for current status data. *Lifetime Data Anal.* **4**, 29–50.
van de Geer, S. A. (1990). Estimating a regression function. *Ann. Statist.* **18**, 907–924.
van der Vaart, A. & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer, New York.
Wellner, J. (2003). Gaussian white noise models: some results for monotone functions. *Crossing boundaries: statistical essays in honor of Jack Hall*. (eds J. E. Kolassa & D. Oakes). IMS Lecture Notes – Monograph Series, Vol. 43, pp. 87–104.
Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika* **89**, 39–48.

Received November 2004, in final form December 2005

Moulinath Banerjee, Department of Statistics, 439, West Hall, 1085, S. University, Ann Arbor, MI 48109, USA.
E-mail: moulib@umich.edu

Appendix

Proof of theorem 2

The proof of this theorem relies on extensive use of 'switching relationships' which allow us to translate the behaviour 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 behaviour of the slope process can then be studied in terms of the limiting behaviour of the minimizer of this stochastic process by applying argmin continuous mapping theorems. Switching relationships on the limit process then allow interpretation of the behaviour 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 (ψ_n, ϕ_n) converges to (ψ, ϕ) pointwise, we can conclude 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 & 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

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

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:

$$\tilde{\Lambda}(z) \le a \Leftrightarrow \mathrm{argmin}_{r \ge 0} [B_n(r) - aG_n(r)] \ge Z_z \tag{18}$$

where Z_z is the largest $Z_{(i)}$ 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_is. 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

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

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

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

where the second step in the above display follows from the first on using (18), and $V_n(r) = B_n(r) - \Lambda_0(z_0) G_n(r)$. Thus,

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

where

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

and

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

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 $L(h) \equiv \tilde{a} W(h) + \tilde{b} h^2 -$

$xC(z_0)h$. Here $\tilde{a}=\sqrt{C(z_0)}$, $\tilde{b}=\lambda_0(z_0)\,C(z_0)/2$ 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

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

to $\tilde{a}W(h)+\tilde{b}h^2$ [by arguments similar to those in lemma 2.3 of Banerjee (2005)] along with the fact that $\sup_{h\in[-M,\,M]}|\tilde{\Lambda}(z_0+hn^{-1/3})-\Lambda_0(z_0)|=O_p(n^{-1/3})$ which entails that $\sup_{h\in[-K,\,K]}|P_{n,\Lambda_0}(h)-\mathbb{M}_n(h)|\to_p 0$, 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 $\operatorname{argmin}_h(\mathbb{M}_n(h)-x\,\mathbb{G}_n(h))$ to $\operatorname{argmin}_h L(h)$ 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 & Wellner (1996)]. The computations are tedious but straightforward and skipped here. For a flavour of the key steps involved in establishing tightness, we refer the reader to section 3.2.3 of van der Vaart & Wellner (1996) and in particular example 3.2.15 (current status data) which is naturally related to binary regression and pp. 212–216 of Banerjee (2000).

It follows that

$$\begin{aligned}&\lim_{n\to\infty} P\,(n^{1/3}\,(\tilde{\Lambda}(z_0+h_0n^{-1/3})-\Lambda_0(z_0))\le x)\\&\quad=P\,(\operatorname{argmin}_{\mathbb{R}}\,\tilde{a}W(h)+\tilde{b}h^2-xC(z_0)\,h\ge h_0).\end{aligned}\tag{19}$$

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

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

with probability 1. Therefore,

$$\lim_{n\to\infty} P\,(n^{1/3}\,(\tilde{\Lambda}(z_0+h_0n^{-1/3})-\Lambda_0(z_0))\le x)=P\,(g_{\tilde{a},\tilde{b}}(h_0)<x\,C(z_0)).$$

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 1) our proof is complete.

Proof of theorem 3

The likelihood ratio statistic of interest can be written as

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

It will follow from theorem 1 that

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

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 (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

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

by Taylor-expanding $\phi_i(t)$ around $\Lambda_0(z_0)$. Now consider $T_{n,2}$. Once again, by Taylor expansion, we have

$$T_{n,2} = \sum_{i \in J_n} \frac{1}{2}\phi_i''(\tilde{\Lambda}(z_i))[(\tilde{\Lambda}(z_{(i)}) - \tilde{\Lambda}_0(z_0)]^2 - \sum_{i \in J_n} \frac{1}{2}\phi_i''(\tilde{\Lambda}_0(Z_{(i)}))[\tilde{\Lambda}_0(Z_{(i)}) - \tilde{\Lambda}_0(z_0)]^2 + o_p(1). \tag{20}$$

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. By Taylor expansion again,

$$S_2 = -\sum_{i \in J_n} \phi_i''(\tilde{\Lambda}_0(Z_{(i)}))\,\left[\tilde{\Lambda}_0(Z_{(i)}) - \frac{\phi_i'(\tilde{\Lambda}_0(Z_{(i)}))}{\phi_i''(\tilde{\Lambda}_0(Z_i))} - \Lambda_0(z_0)\right]\,(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0)) + o_p(1).$$

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,

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

using the following observation: if B' is a level block for $\tilde{\Lambda}_0$ contained in J_n with level value w^0, then

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

provided $w^{(0)} \neq \theta_0$. This is a direct consequence of the representation (13). Conclude that $S_2 + o_p(1) = -\sum_{i \in J_n} \phi_i''(\tilde{\Lambda}_0(Z_{(i)}))(\tilde{\Lambda}_0(Z_{(i)}) - \Lambda_0(z_0))^2$. It is similarly established (using (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

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

Now, on using (20), we get

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

whence

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

Invoking the explicit representation for $\phi_i''(\Lambda_0(Z_{(i)}))$, some algebra shows that

$$C_n = n^{1/3} (\mathbb{P}_n - P)\Psi_n(\delta, z, x) + n^{1/3} P\Psi_n(\delta, z, x) + o_p(1)$$

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

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

D_n denoting the set where $\tilde{\Lambda}$ and $\tilde{\Lambda}_0$ differ. 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 Pg 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 $[z_0 - Mn^{-1/3}, z_0 + Mn^{-1/3}]$ with arbitrarily 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 $\Psi_n(\delta, z, x)$ 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

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

Some algebra, along with the facts that $\tilde{D}_n \equiv n^{1/3}(D_n - z_0)$ is eventually contained with arbitrarily high probability in a compact set and the boundedness in probability of the processes $U_n(h)$ and $V_n(h)$ for h in compacts, yields the representation:

$$n^{1/3} P\Psi_n(\delta, z, x) = \int_{\tilde{D}_n} C(z_0)(U_n^2(h) - V_n^2(h))\,dh + o_p(1).$$

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

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

and the fact that

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

follows as a direct application of lemma 1 followed by the change of variable theorem from calculus.

It remains to show that

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

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

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

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 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 $O_p(1)$.