

Robust Differentiable Functionals for the Additive Hazards Model

Enrique E. Álvarez¹, Julieta Ferrario²

¹Schools of Economics and Engineering, Universidad Nacional de La Plata y CONICET, La Plata, Argentina

²Department of Mathematics, Universidad Nacional de La Plata, Buenos Aires, Argentina

Email: enriqueealvarez@fibertel.com.ar, jferrario@mate.unlp.edu.ar

Received 10 September 2015; accepted 27 October 2015; published 30 October 2015

Copyright © 2015 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

In this article, we present a new family of estimators for the regression parameter β in the *Additive Hazards Model* which represents a gain in robustness not only against outliers but also against unspecific contamination schemes. They are consistent and asymptotically normal and furthermore, and they have a nonzero breakdown point. In Survival Analysis, the *Additive Hazards Model* proposes a hazard function of the form $\lambda(t) = \lambda_0(t) + \beta'z$, where $\lambda_0(t)$ is a common nonparametric baseline hazard function and z is a vector of independent variables. For this model, the seminal work of Lin and Ying (1994) develops an estimator for the regression parameter β which is asymptotically normal and highly efficient. However, a potential drawback of that classical estimator is that it is very sensitive to outliers. In an attempt to gain robustness, Álvarez and Ferrario (2013) introduced a family of estimators for β which were still highly efficient and asymptotically normal, but they also had bounded influence functions. Those estimators, which are developed using classical Counting Processes methodology, still retain the drawback of having a zero breakdown point.

Keywords

Robust Estimation, *Additive Hazards Model*, Survival Analysis

1. Introduction

In Survival Analysis, a main goal is how to model a random variable T^* which is nonnegative and typically continuous and represents the waiting time until some events. A common method for collection of survival-type data consists in deciding on an observation window $[0, \tau]$ over which n individuals are followed. Naturally, some events may take longer to occur than the window length τ ; also, some individuals can be lost from the

sample due to different reasons (such as changing hospitals in clinical studies). In those cases, instead of an event time a censoring time is observed. In that manner, at the end of the observation window, the researcher ends up with a sample of triplets (T_i, Z_i, Δ_i) where $T_i = \min(T_i^*, C_i)$ represents either the true duration or the censoring time, $\Delta_i = 1(T_i = T_i^*)$ is the indicator that the observed time is uncensored, and Z_i is a vector of individual covariates. Statistical models for that type of data are the main goal of the branch of statistics called Survival Analysis, and the relevant literature is by now enormous. Some classical textbook-long treatises were Kalbfleish and Prentice (1980) [1], Fleming and Harrington (1991) [2], Andersen *et al.* (1993) [3], Aalen *et al.* (2008) and [4], among others.

A particular type of survival models with great appeal among practitioners focuses on the so-called *hazard function* $\lambda(t)$, which intuitively measures the instantaneous risk of the occurrence of the event at any given moment in time. While the most widespread model for the hazard function was the semiparametric *Multiplicative Hazards Model* due to Cox (1972) [5], a popular alternative for datasets without proportionality of hazards was the *Additive Hazards Model* (AHM) presented by Aalen (1980) [6]. With time-fixed covariates, the latter proposes that $\lambda(t, z) = \lambda_0(t) + \beta'z$, where β is a vector of p nonnegative parameters. An estimation method for β and the nonparametric baseline function λ_0 for this model were first described in a seminal article by Lin and Ying (1994). They proposed an estimating equation for the Euclidean parameter β which was independent of $\lambda_0(t)$ and which had the additional benefit of yielding an estimate in closed form, in addition to being consistent and asymptotically normal. Its drawback, however, lies in the sensitivity to outliers.

Within the Cox model, the potential harmful effects of outliers were commented by Kalbfleisch and Prentice (1980, ch. 5) [1] and Bednarski (1989) [7]. Robust alternatives were first introduced by Sasieni (1993a, 1993b) [8] [9] by essentially modifying the Cox's partial likelihood score function introducing weight functions. Along the same line, important work had been developed by Bednarski (1993) [10], who proposed estimators that were consistent and efficient not only at the model but also on small contaminated neighbourhoods. His estimators had the advantage of being Fréchet differentiable for a wide class of weight functions.

As for the *Additive Hazards Model*, the proposal of robust alternatives has received much less attention in the literature. In Álvarez and Ferrario (2013) [11], we introduced a family of estimators for the Euclidean parameter β in the AHM by introducing weights in Lin and Yings' estimating equation. The weight functions are carefully chosen so that the estimators remain Fisher-consistent and asymptotically normal, but they additionally represent a gain in robustness at the price of a modest loss in efficiency. The proposed family of estimators exhibits a gain in robustness against the classical LY (Lin and Yings) estimator because they have a bounded influence function. That type of robustness is qualitative in nature, and it means intuitively that an estimation method is able to tolerate a very small proportion of extreme values. To see this, recall that the Influence Function is heuristically the population version of the so called *Sensitivity Curve*.

$$SC_n(t_0, z_0, \delta_0) := \frac{\hat{\beta}_n^*(t_0, z_0, \delta_0) - \hat{\beta}_n}{1/n}, \quad (1)$$

where $\hat{\beta}_n$ is the estimator from the pure (noncontaminated) sample and $\hat{\beta}_n^*(t_0, z_0, \delta_0)$ is the estimator in a contaminated sample where one random observation has been replaced by the triplet (t_0, z_0, δ_0) . *i.e.* the population contamination model for the triplet $(T, Z, \Delta) \sim H^c$ is of the form

$$H^c := (1 - \varepsilon)H + \varepsilon Q_{(t_0, z_0, \delta_0)}, \quad (2)$$

where H is the noncontaminated distribution that belongs to the additive hazards family and Q represents a point mass at its argument. For the practitioner, estimators with bounded influence functions are of interest when (s) he seeks a guard against a very small proportion of outliers.

Appart from the fact that the contamination scheme above is very specific, a further drawback of the estimators presented in Álvarez and Ferrario (2013) [11] is that they have a zero breakdown point. Heuristically this means that just a small proportion of contamination, strategically located, is sufficient to drive the estimators nonsensical. Different notions and measures of robustness and their implications are developed in many classical books, such as Maronna, Martin and Yohai (2006) [12], Huber and Ronchetti (2009) [13] and Hampel *et al.* (1986) [14].

In this article we propose a new family of robust estimators for the additive hazards model in a manner similar to Bednarski (1993) [10]. This is, we start from Lin and Yings' estimating equation and modify it by introducing

appropriate weight functions that retain consistency and asymptotic normality while improving robustness, in the sense that the resulting estimating functionals are Fréchet differentiable about small neighborhoods of the true model. That type of differentiability entails three important consequences: 1) that the proposed family of estimators has bounded influence functions; 2) that they have a strictly positive (nonzero) breakdown point; and 3) that consistency and asymptotic normality hold about small neighbourhoods of the true model for generic contamination schemes, *i.e.* our family of estimators resists not only the outlier-type contamination presented in (2) but also small deviations in the structure of the model itself. For instance, one could contaminate with model a which does not have additive hazards, or with a model in which T and Δ are dependent, even conditional on Z . That makes the proposed family of estimators attractive from the practitioner point of view, as a backup against model misspecification.

The advantage of the estimators we present in this paper over previous proposals arises whenever a dataset contains outliers. When a sample is contaminated by unusual observations, the classical estimator (Ling and Yings) rapidly becomes nonsensical (in that its value drifts away towards zero or infinity). The estimators in Álvarez and Ferrario [11] on the other hand, while they resist contamination by large times or large values of the covariates, they exhibit no advantage against more involved types of contamination. Here we develop a family of estimators that resist arbitrary contamination schemes. This paper is organized as follows, in section we introduce the estimating method and we construct explicitly the *Additive Hazards Family* of distribution functions for survival data. In subsection we prove that our estimators are Fréchet differentiable. That entails asymptotic normality not only at true distributions in the additive hazards family, but also under contiguous alternatives. In order to assess the performance of the proposed method in small samples, section contains a small simulation study which serves two purposes: 1) it illustrates the improvement of our proposed estimators from the robustness point of view against the classical counterparts; and 2) it exhibits a non-zero breakdown point which is apparently fairly high. A simulation approach to the breakdown point is important because it is not feasible to compute it analytically. That is in part because the calculations involve are formidable, as they involve identifying the worst possible contaminating distribution. But more importantly, it is because the breakdown point depends on the joint distribution of the triplet (T, Z, Δ) , which is only specified semi-parametrically in the *Additive Hazards Model*, *i.e.* even if available, the closed-form expression for the theoretical breakdown point would depend on unknown quantities. We have written computing code for the method proposed in this article in the form of R-scripts, which is available from the authors upon request. All the proofs are presented as succinctly as possible in the Appendix; further calculations can be found in Julieta Ferrario’s PhD dissertation [15].

2. Robust Differentiable Estimators

Let $\{N_i(t) : t \geq 0\}$ be the counting process which records the occurrence of the event for individual i , this is $N_i(t) = 1(T_i^* \leq t)$, so that it jumps from zero to unity at the random time T_i^* ; and let $\{Y_i(t) : t \geq 0\}$ be the so-called *at risk process* defined by $Y_i(t) := 1(T_i^* \wedge C_i > t)$, which denotes that by time t neither the terminal event nor censoring has occurred for individual i , so that (s) he is still at risk. For the *Additive Hazards Model*, Lin and Ying (1994) [16] proposed the estimating equation

$$U_n(\hat{\beta}) := \sum_{i=1}^n \int_0^\infty \left\{ [z_i - \bar{z}_n(t)] dN_i(t) - Y_i(t) [z_i - \bar{z}_n(t)]^{\otimes 2} \hat{\beta}_n dt \right\} = 0, \tag{3}$$

where for a column vector v , we denote the matrix $v^{\otimes 2} := vv'$, and we define the process $\bar{z}_n(t) := \sum 1(t_i \geq t) z_i / \sum 1(t_i \geq t)$. This yields an estimator in closed form

$$\hat{\beta}_n = \left[\sum_{i=1}^n \int_0^\infty Y_i(t) [z_i - \bar{z}_n(t)]^{\otimes 2} dt \right]^{-1} \left[\sum_{i=1}^n \int_0^\infty [z_i - \bar{z}_n(t)] dN_i(t) \right]. \tag{4}$$

Using classical Counting Process theory, Lin and Ying prove that their $\hat{\beta}_n$ is consistent and asymptotically normal, and they provide a formula for the estimation of the asymptotic variance.

In order to propose a Fréchet differentiable alternative to the classical (Lin and Ying’s) estimator we need to express the estimator as a functional of the joint empirical distribution function and we need to make explicit the structure of the *Additive Hazard Family* of distributions. We pursue this as follows.

2.1. Construction of the Additive Hazards Family

Event times: Let T^* represent the event time, which could be unobserved when censored. Given the covariate values $\{Z = z\}$, T^* has a hazard (conditional) function $\lambda_*(t, z) = \lambda_0(t) + \beta'z$. Correspondingly, the Survival function is $S_*(t) = S_0(t)\exp(-\beta'zt)$, and the density function is $f_*(t) = f_0(t)\exp(-\beta'zt) + S_0(t)\beta'z\exp(-\beta'zt)$.

Covariates: The covariates $Z \sim F_Z$ are nonnegative and nondegenerate.

Censoring: Conditional on Z , censoring and event times are independent, i.e. $F_{T^*,C,Z} = F_{T^*|Z} F_{C|Z} F_Z$.

Observed times: Due to censoring, the observed times are $T := T^* \wedge C$. Conditional on $\{Z = z\}$ their survival function is $P(T > t | Z = z) = S_*(t)S_c(t)$. This entails

$f_{T|Z}(t, z) = S_0(t)S_c(t)\exp(-\beta'zt)[\lambda_0(t) + \lambda_c(t) + \beta'z]$. Therefore,

the joint density of T and Z is

$$h_{T,Z}(t, z) := S_0(t)S_c(t)\exp(-\beta'zt)[\lambda_0(t) + \lambda_c(t) + \beta'z]f_Z(z). \tag{5}$$

We now develop the joint bivariate distribution function $H(t, \Delta = 1)$ that corresponds only to the noncensored times.

Censoring indicator: Let $\Delta = 1(T^* \leq C)$ be the indicator of a noncensored observation. Consider

$$\begin{aligned} P(T > t; \Delta = 1 | Z = z) &= P(T > t | Z = z) - P(T > t; \Delta = 0 | Z = z) \\ &= S_*(t)S_c(t) - E\left[P(t < C < T^* | Z = z) | T^*\right] \\ &= S_*(t)S_c(t) - \int_t^\infty [S_c(t) - S_c(u)]f_*(u)du. \end{aligned}$$

Thus taking the derivative with respect to t we obtain

$$h_{T,Z,\Delta}(t, z, \Delta = 1) := S_0(t)S_c(t)[\lambda_0(t) + \beta'z]\exp(-\beta'zt)f_Z(z). \tag{6}$$

We define that a cumulative joint distribution function $H(t, z, \delta)$ for the triplet (t, Z, Δ) is a member of the Additive Hazards Family if 1) it has a joint bivariate distribution function for (t, z) as given by $H(t, z)$ in Equation (5) and 2) if it has a restricted distribution function $H(t, z, \Delta = 1)$ as given in Equation (6). It is noteworthy that the additive family \mathcal{A} is semiparametric, as it is indexed by $\beta \in \mathbb{R}^{p^+}$, but also the arbitrary survival functions $S_0(\cdot)$, $S_c(\cdot)$ and $S_z(\cdot)$.

Now we express the classical estimator $\hat{\beta}_n$ as a (nonlinear) functional of the joint empirical distribution function of (T, Z, Δ) .

$$\hat{\beta}_n = \left[\iint \left(\int_0^t [z - \bar{z}_n(u)]^{\otimes 2} du \right) d\mathbb{H}_n(t, z) \right]^{-1} \left[\iint [z - \bar{z}_n(t)] d\mathbb{H}_n(t, z, \Delta = 1) \right], \tag{7}$$

where we introduce the process $\bar{z}_n(u) = \left(\iint \mathbf{1}(u \leq t) z d\mathbb{H}_n(t, z) \right) / \left(\iint \mathbf{1}(u \leq t) d\mathbb{H}_n(t, z) \right)$ and the \mathbb{H}_n 's are the empirical distributions.

In Alvarez and Ferrario (2013) [11] we illustrated that the classical estimator is very sensitive to outliers and we have shown that its influence function is unbounded. In this article we propose an alternative family of estimators which is robust not only against outliers but also against unspecific contamination, in that the defining functional is not only continuous but also Fréchet differentiable. This entails a nonzero breakdown point and bounded influence curve. As a reference, the implication and uses of Fréchet differentiable statistical functionals in Asymptotic Statistics and Robust Statistics are thoroughly presented in Bednarski (1991) [17].

2.2. Fréchet Differentiability

In order to define contamination ε -neighborhoods, let $H \in \mathcal{A}$ be a distribution in the additive hazards family with true parameter $\beta_0 \in \mathbb{R}^{p^+}$ and let G be some contaminating distribution for the triplet (T, Z, Δ) . We say that G is in a neighborhood $\mathcal{B}(H, \varepsilon)$ of H if $\|G - H\|_\infty := \sup_{(t,z,\delta)} |G(t, z, \delta) - H(t, z, \delta)| \leq \varepsilon$. Note that if G is a point mass at some triplet (t_0, z_0, δ_0) this contamination scheme corresponds to what are most usually called

outliers. Our formulation is more general in that an arbitrary measure G may introduce model misspecification also from a disruption of the additive hazards property, of the conditional independence of the event and the censoring times given the covariates, or any other feature of the additive hazards family.

We propose here an estimating equation by introducing weight functions in the classical formulation, *i.e.*

$$L(\mathbb{H}_n, \beta, W) := \left[\iint W(t, z) [z - z_{\mathbb{H}_n}(t)] d\mathbb{H}_n(t, z, \Delta = 1) \right] - \left[\iint W(t, z) \left(\int_0^t [z - z_{\mathbb{H}_n}(u)] du \right) z' d\mathbb{H}_n(t, z) \right] \beta = 0, \tag{8}$$

where $W : \mathbb{R}^+ \times \mathbb{R}^{p^+} \mapsto \mathbb{R}^+$ is a bivariate weight function in a set \mathcal{W} , the properties of which will be enumerated below. Also,

$$z_{\mathbb{H}_n}(t) = \frac{\iint I(u \geq t) W(t, x) x d\mathbb{H}_n(u, x)}{\iint I(u \geq t) W(t, x) d\mathbb{H}_n(u, x)}. \tag{9}$$

Naturally, in the special case where $W(u, z) = 1$ we obtain the classical (Lin and Yings') estimator. Instead, if $W(u, z) = w_1(u)w_2(z)$ and both factors are either deterministic or predictable stochastic processes, this gives the estimators proposed in Álvarez and Ferrario (2013) [11]. In that article, the choice of weight function led to estimators in closed form, but more importantly, it made possible for all the properties and proofs to be developed using Counting Process Martingale Theory. We depart from that treatment in this article, as we specify the weight functions differently, *i.e.* we do not seek here predictability as stochastic processes, but Fréchet differentiable functionals.

Let us denote by $\beta(G)$ the solution of $L(G, \beta(G), W) = 0$. In order to find a linear approximation for $\beta(G) - \beta_0$ we need to choose some type of differentiability. We opt here for differentiability in the sense of Fréchet (or *strong differentiability*). This type of differentiability is stronger than continuity in that it implies the existence of a linear functional approximation, *i.e.* for any $H \in \mathcal{A}$

$$\lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \sup_{\substack{G \in \mathcal{B}(H, \varepsilon) \\ W \in \mathcal{W}}} \left\| [\beta(G) - \beta_0] - \text{FD}(G - H) \right\|_2 = 0, \tag{10}$$

where FD is a linear functional called “Fréchet derivative”. Notice that we opt here for a uniform type of differentiability over \mathcal{W} . This is important in order to allow for instance adaptive choice of the weight function, such those based on preliminary estimators of β . In contrast with this generality, data-dependent choices were explicitly excluded in our first proposal of estimators in [11], for that mechanism would automatically destroy the predictability of the stochastic weight processes.

In order to avoid excessive notation we will in the sequel develop the proofs without censoring. Let $\beta_0 \in B$, where B denotes a open bounded subset of \mathbb{R}^{p^+} and take a member in the additive hazards family $H \in \mathcal{A}$ corresponding to the true value of the parameter β_0 . Take further a set of functions $\mathcal{W}^* := \{W^* : \mathbb{R}^+ \times \mathbb{R}^{p^+} \mapsto \mathbb{R}^+\}$, and a real function $W_0 : \mathbb{R}^+ \mapsto \mathbb{R}^+$ which is continuous and has a bounded support $K := [a, b]$, for some $0 < a < b < \infty$. With the above, we define the family of weight functions \mathcal{W} with which we work in this article by $\mathcal{W} := \{W_0 W^* : W^* \in \mathcal{W}^*\}$.

In order that our family of estimators become Fréchet differentiable we will need the following assumptions.

Assumptions

A1) For all $W^* \in \mathcal{W}^*$, $u \in K$, the integral $\iint W^*(u, z) I(t \geq u) dH(t, z) > \varepsilon$, for some $\varepsilon > 0$.

A2) All the functions in \mathcal{W}^* vanish outside some bounded set, are absolutely continuous and have jointly bounded variation. The set \mathcal{W}^* is compact with respect to the supremum norm.

Assumptions A1) and A2) ensure differentiability. The compactness assumption in A2) is needed to allow possibly adaptive choices of W based on some preliminary estimate of β (*i.e.* in a data-dependent manner). The assumption of jointly bounded variation is needed in order to obtain uniform Fréchet differentiability over \mathcal{W} .

We seek now a linear approximation of $\beta(G) - \beta_0$ in terms of $G - H$. Firstly, consider

$$\begin{aligned} & L(G, \beta, W) - L(H, \beta_0, W) \\ &= [L(G, \beta, W) - L(H, \beta, W)] + [L(H, \beta, W) - L(H, \beta_0, W)]. \end{aligned} \tag{11}$$

For the first difference in functionals above, the following Lemma gives a linear approximation:

Lemma 1. Under assumptions A1) and A2),

$$\sup_{\substack{\beta \in B \\ W \in \mathcal{W}}} \|L(G, \beta, W) - L(H, \beta, W) - \text{Lin}_{W, \beta}(G - H)\|_2 = o(\|G - H\|_\infty^2),$$

where

$$\begin{aligned} \text{Lin}_{W, \beta}(G - H) &= \iint W(t, z) [z - z_H(t)] d(G - H)(t, z) \\ &\quad - \left\{ \iint W(t, z) \left(\int_0^t [z - z_H(u)] du \right) z' d(G - H)(t, z) \right\} \beta_0 \\ &\quad + \iint W(t, z) [z_H(t) - z_G(t)] dH(t, z) \\ &\quad - \left\{ \iint W(t, z) \left(\int_0^t [z_H(u) - z_G(u)] du \right) z' dH(t, z) \right\} \beta. \end{aligned} \tag{12}$$

Moreover, $\text{Lin}_{W, \beta_0}(H) = 0$.

As for the second difference in (11) we have:

Lemma 2. For any $W \in \mathcal{W}$ and $\beta \in B$,

$$L(H, \beta, W) - L(H, \beta_0, W) = \mathcal{L}(H, W)(\beta - \beta_0), \tag{13}$$

where $\mathcal{L}(H, W)$ is the matrix of partial derivatives

$$\mathcal{L}(H, W) := \frac{\partial L(H, \beta, W)}{\partial \beta} = - \left[\iint W(t, z) \left(\int_0^t [z - z_H(u)] du \right) z' dH(t, z) \right]. \tag{14}$$

Further, the following result gives a bound of $\text{Lin}_{W, \beta}(G - H)$ in terms of $(G - H)$:

Lemma 3. Under assumptions A1) and A2) there are constants $c > 0$ and $\varepsilon > 0$ such that for all $W \in \mathcal{W}$ and $\beta \in B$, if $G \in \mathcal{B}(H, \varepsilon)$ then

$$\|\text{Lin}_{W, \beta}(G - H)\|_2 \leq c \|G - H\|_\infty. \tag{15}$$

At this point, for further results we need to add another assumption that guarantees the existence of the inverse of $\mathcal{L}(H, W)$ throughout, namely.

A3) There is a pair of constants $a_1, a_0 > 0$, so that for all $W \in \mathcal{W}$, the determinant is bounded, i.e. $a_0 \leq |\det \{(H, W)\}| \leq a_1$.

Thus, the consistency of the estimator in a neighborhood of \mathcal{A} is given by the following theorem:

Theorem 1. Let the family of functions \mathcal{W} satisfy Assumptions A1) though A3). Then there exist $\varepsilon > 0$ and $M > 0$ so that if $G \in \mathcal{B}(H, \varepsilon)$, the equation $L(G, \beta, W) = 0$ has a solution in the ball $\|\beta - \beta_0\|_2 < M \|G - H\|_\infty$ for all $W \in \mathcal{W}$.

Moreover, Fréchet differentiability is asserted as follows:

Theorem 2. Let $\beta(G, W)$ denote a solution of the Equation $L(G, \beta, W) = 0$. If the class of functions \mathcal{W} satisfy Assumptions A1) through A3) then:

$$\lim_{c \rightarrow 0} \sup_{\substack{W \in \mathcal{W} \\ G \in \mathcal{B}(H, c)}} \frac{1}{c} \left\| \left[\beta(G, W) - \beta_0 \right] + \mathcal{L}^{-1}(H, W) \text{Lin}_{W, \beta_0}(G - H) \right\|_2 = 0.$$

This implies that the Fréchet derivative of $\beta(H, W)$ at β_0 towards $(G - H)$ is $FD(G - H) := \mathcal{L}^{-1}(H, W) \text{Lin}_{W, \beta_0}(G - H)$.

In the following theorem we investigate convergence in distribution under contiguous alternatives to some distribution in the additive hazards family $H \in \mathcal{A}$ with true parameter β_0 :

Theorem 3. Let \mathbb{H}_n be the empirical distribution of a sample of size n from a distribution $G \in \mathcal{B}(H, c_0/\sqrt{n})$ for some constant $c_0 > 0$. Assume the class of functions \mathcal{W} satisfies A1) through A3). Then for all $\varepsilon > 0$ and $\delta > 0$ there exists n_0 so that for all $n \geq n_0$, $G \in \mathcal{B}(H, c_0/\sqrt{n})$, and $W \in \mathcal{W}$, we have

$$G^n \left(\sup_{W \in \mathcal{W}} \sqrt{n} \left\| \beta(\mathbb{H}_n, W) - \beta_0 - \iint \mathcal{IF} d\mathbb{H}_n \right\|_2 > \varepsilon \right) < \delta,$$

where $\mathcal{IF}(x, \beta_0, W) := -\mathcal{L}^{-1}(H, W) \text{Lin}_{W, \beta_0}(\delta_x - H)$ and δ_x is a point mass at $x := (t, z, \Delta)$.

The result above implies that asymptotic normality holds not only under the true model but also under contiguous alternatives.

3. Simulations

In this section, we evaluate the performance of our proposed family of estimators via simulations. Specifically, we carry out three simulation experiments choosing for simplicity a single covariate ($p=1$) and a true additive hazard function $\lambda_*(t, z) = 1 + 0.5z$, with $Z \sim \chi_1^2$ and, in order to avoid censoring $C=1000$. In all cases, in the estimating Equation (8), we opt for $W(t_i, z_i) = 1 - \min(1, z_i/M)$, where M is the lower $m\%$ percentile of the z 's.

In the first simulation we study the behavior of our estimator, denoted ‘‘RD’’ (Robust Differentiable) for increasing sample sizes. We take $n = 50, 200, 500, 1000$ and $10,000$. In the function $W(t_i, z_i)$, we select $m = 90, 95$ and 99 . **Table 1** compares our estimators with the classical ones, denoted ‘‘LY’’ (Lin and Yings’). We perform 100 replicates and average out the results. We observe in all cases a very good performance of the of the robust estimators at very small price in increased standard error, which decreases as m or n increase.

In the second simulation, we do a comparison among the classical estimator (LY), the bounded-influence-function (BIF) estimators proposed in Álvarez and Ferrario (2013) [11], and the ones proposed in this paper (RD). This is done under outlier-type contamination, where an increasing percentage of the sample was replaced by a large covariate value equal to 10. We take 100 replicates for a sample size of $n = 200$. In **Table 2**, we show the results of this experiment. For the BIF estimators we take the weight function $W(t_i, z_i) = \min(1, a\tilde{z}/z_i)^2$ where

Table 1. Classical vs. robust differentiable estimators in pure samples.

Estimator		LY	RD		
n			$m = 90$	$m = 95$	$m = 99$
50	coef.	0.59	0.51	0.52	0.54
	(s.e.)	(0.32)	(0.50)	(0.43)	(0.34)
200	coef.	0.51	0.52	0.52	0.51
	(s.e.)	(0.15)	(0.20)	(0.17)	(0.13)
500	coef.	0.51	0.51	0.51	0.51
	(s.e.)	(0.09)	(0.13)	(0.10)	(0.09)
1000	coef.	0.51	0.49	0.49	0.50
	(s.e.)	(0.07)	(0.09)	(0.08)	(0.06)
10,000	coef.	0.50	0.51	0.51	0.50
	(s.e.)	(0.02)	(0.03)	(0.02)	(0.02)

Table 2. Comparison of estimators with outliers.

%	Estimators	Pure Sample		With Outliers	
		LY	LY	BIF	RD
0	coef.	0.55	0.55	0.54	0.53
	(s.e.)	(0.19)	(0.19)	(0.16)	(0.22)
0.5	coef.	0.54	0.40	0.46	0.52
	(s.e.)	(0.18)	(0.23)	(0.16)	(0.23)
5	coef.	0.53	0.11	0.22	0.54
	(s.e.)	(0.14)	(0.56)	(0.41)	(0.18)
15	coef.	0.50	0.040	0.08	0.52
	(s.e.)	(0.15)	(0.66)	(0.59)	(0.16)
25	coef.	0.50	0.02	0.05	0.51
	(s.e.)	(0.14)	(0.67)	(0.64)	(0.16)

$a = 2.3026$ and $\tilde{s} = \text{median}\{z_1, \dots, z_n\} / \log(2)$. With that weight function, 90% of the observations were unmodified if Z had an exponential distribution (e.g. Álvarez and Ferrario 2013 [11]). For the RD estimators we opt for $m = 90$. We observe that while both the Classical and the BIF estimators appear to break down rather fast, the robust estimator behaves very well. It is worth to emphasize that the type of contamination here is not the *worst possible* one, so that there is no reason to expect the good performance of the RD estimators to extend to other contamination schemes.

Lastly, we carry out a third simulation experiment in order to detect what the breakdown points of the RD estimators may be under a different type of model departure. As a model for the contaminating distribution, we chose point masses on the line $T = 4 + 0.25Z$. This artificially introduces outliers in the sample as shown in Figure 1, where the red observations in the first plot are replaced by the blue-colored points shown in the second plot. This graphical illustration shows that the contaminating distribution is very different from the proposed *Additive Hazards Model* and it may thus have the ability to severely affect the estimates. If, instead, we have contaminated by large values of z (high leverage) or large values of t (outliers), keeping the other variable unchanged, the potential harmful effects are greatly diminished. This is because we can see in the first (uncontaminated) plot many of such points. The simulation results of 200 replicates are shown in Table 3, for

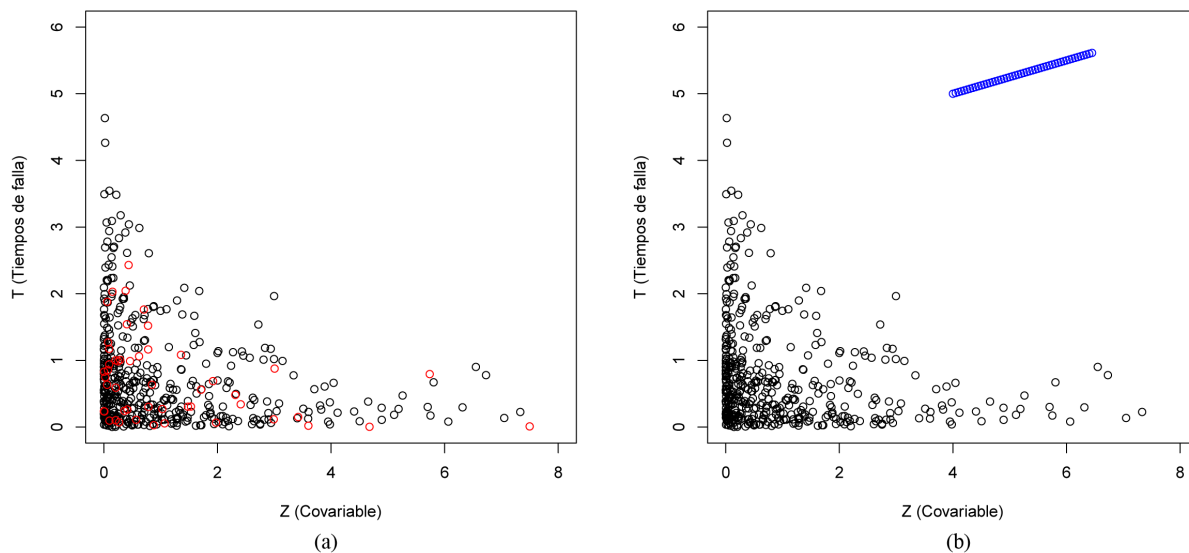


Figure 1. Pure vs. contaminated sample. (a) Pure sample (b) Contaminated sample.

Table 3. Comparison of estimators under model contamination.

%		Pure Sample		Contaminated Sample	
		$m = 70$	$m = 90$	$m = 70$	$m = 90$
0.2	coef.	0.55	0.52	0.55	0.52
	(s.e.)	(0.28)	(0.14)	(0.28)	(0.14)
3	coef.	0.50	0.51	0.50	0.51
	(s.e.)	(0.27)	(0.14)	(0.25)	(0.13)
10	coef.	0.50	0.52	0.51	0.21
	(s.e.)	(0.29)	(0.14)	(0.23)	(0.42)
20	coef.	0.50	0.50	0.50	-0.13
	(s.e.)	(0.27)	(0.13)	(0.15)	(0.89)
25	coef.	0.48	0.51	0.51	-0.16
	(s.e.)	(0.27)	(0.14)	(0.15)	(0.93)
30	coef.	0.51	0.53	0.22	-0.17
	(s.e.)	(0.27)	(0.14)	(0.40)	(0.94)

sample a size $n = 500$. We observe that the breakdown point seems to be about 30% for $m = 70$ and at about 10% for $m = 10$. This is intuitive as the constant m regulates the trimming, given our choice of the weight function.

Intuitively, the finite sample breakdown point of an estimator is the largest proportion of contaminated observations, and a method can resist before the estimates become nonsensical, which usually means that the estimate drifts away towards zero or infinity, or in general towards the boundaries of a parameter space. Equivalently, its functional version is called the asymptotic breakdown point and it measures the largest proportion of contamination. A statistical functional could tolerate before becoming nonsensical in the same sense (e.g. Maronna, Martin and Yohai 2006 [12] for formal definitions). It is noteworthy that either in its finite sample or in its asymptotic version, calculating a breakdown point requires identifying the worst possible type of contamination. This would depend on the joint distribution of the triplet (T, Z, Δ) , which is only partially specified in the *Additive Hazards Model*. Therefore, the breakdown point cannot be calculated explicitly. Nor is it feasible to provide reasonable bounds. For that reason, it is not possible to give a numeric value for the breakdown point and its assessment via simulations becomes illustrative. An extensive investigation of the breakdown point under different distributions $H(t, z, \delta)$ and under different weight functions $W(t, z)$ via simulation is a subject of further research.

Acknowledgements

We thank the editor and the referee for their comments. This work has been financed in part by UNLP Grants PPID/X003 and PID/X719. Julieta Ferrario further wishes to thank Tadeus Bednarski for generously sharing otherwise electronically unavailable manuscripts.

References

- [1] Kalbfleisch, J.D. and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data*. Wiley, New York.
- [2] Fleming, T.R. and Harrington, D.P. (1991) *Counting Processes and Survival Analysis*. Wiley, New York.
- [3] Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. Springer-Verlag, New York. <http://dx.doi.org/10.1007/978-1-4612-4348-9>
- [4] Aalen, O.O., Borgan, O. and Gjessing, H.K. (2008) *Survival and Event History Analysis. A Process Point of View*. Springer, New York. <http://dx.doi.org/10.1007/978-0-387-68560-1>
- [5] Cox, D.R. (1972) Regression Models and Life-Tables (with Discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **34**, 187-220.
- [6] Aalen, O.O. (1980) A Model for Nonparametric Regression Analysis of Counting Processes. In: Klonecki, N., Koesh, A. and Rosinski, J., Eds., *Lecture Notes in Statistics*, 2th Edition, Springer, New York, 1-25. http://dx.doi.org/10.1007/978-1-4615-7397-5_1
- [7] Bednarski, T. (1989) On Sensitivity of Cox's Estimator. *Statistics & Decisions*, **7**, 215-228. <http://dx.doi.org/10.1524/strm.1989.7.3.215>
- [8] Sasieni, P. (1993) Maximum Weighted Partial Likelihood Estimators for the Cox Model. *Journal of the American Statistical Association*, **88**, 144-152.
- [9] Sasieni, P. (1993) Some New Estimators for Cox Regression. *Annals of Statistics*, **21**, 1721-1759. <http://dx.doi.org/10.1214/aos/1176349395>
- [10] Bednarski, T. (1993) Robust Estimation in Cox's Regression Model. *Scandinavian Journal of Statistics*, **20**, 213-225.
- [11] Álvarez, E.E. and Ferrario, J. (2013) Robust Estimation in the Additive Hazards Model. *Comm. Statist. Theory Methods*, in press.
- [12] Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006) *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken.
- [13] Huber, P.J. and Ronchetti, E.M. (2009) *Robust Statistics*. Wiley, New York. <http://dx.doi.org/10.1002/9780470434697>
- [14] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.
- [15] Ferrario, J. (2015) *Estimación Robusta en Modelos de Supervivencia con Función de Hazard Aditiva*. PhD Thesis, Universidad Nacional de La Plata, La Plata, Argentina.

- [16] Lin, D.Y. and Ying, Z. (1994) Semiparametric Analysis of the Additive Risk Model. *Biometrika*, **81**, 61-71. <http://dx.doi.org/10.1093/biomet/81.1.61>
- [17] Bednarski, T., Clarke, B.R. and Kolkiewicz, W. (1991) Statistical Expansions and Locally Uniform Fréchet Differentiability. *Journal of the Australian Mathematical Society*, **50**, 88-97. <http://dx.doi.org/10.1017/S1446788700032572>

Appendix

Proof of Lemma 1. Rearranging, $D1 := L(G, \beta, W) - L(H, \beta, W)$ becomes

$$\begin{aligned} D1 &= \iint W(t, z) [z_H(t) - z_G(t)] dG(t, z) + \iint W(t, z) [z - z_H(t)] d(G-H)(t, z) \\ &\quad - \iint W(t, z) \int_0^t [z_H(u) - z_G(u)] du z' dG(t, z) \beta \\ &\quad - \iint W(t, z) \int_0^t [z - z_H(u)] du z' d(G-H)(t, z) \beta. \end{aligned}$$

So that subtracting $Lin_{W, \beta}(G-H)$,

$$\begin{aligned} D1 - Lin_{W, \beta}(G-H) &= \iint W(t, z) [z_H(t) - z_G(t)] d(G-H)(t, z) \\ &\quad - \iint W(t, z) \int_0^t [z_H(u) - z_G(u)] du z' d(G-H)(t, z) \beta. \end{aligned}$$

Let $d(u) := \iint W(u, z) I(t \geq u) dH(t, z)$. Since $W = W_0 W^*$, by A1), $d(u) \geq W_0(u) \varepsilon > 0$. Define further $a(u) := d(u)^{-1} \iint W(u, z) z I(t \geq u) dH(t, z)$, $b(u) := d(u)^{-1} \iint W(u, z) I(t \geq u) d(G-H)(t, z)$, and $c(u) := d(u)^{-1} \iint W(u, z) z I(t \geq u) d(G-H)(t, z)$. Since $b(u) \geq 0$, $|b(u)+1|^{-1} \leq 1$.

To simplify notation, let $I := I(t \geq u)$, $W := W(u, z)$, $dH := dH(t, z)$, $dG := dG(t, z)$ and further $d := d(u)$, $a := a(u)$, $b := b(u)$, $c := c(u)$. Note that $z_H(u) = a$ and $z_G(u) = (a+c)/(b+1)$, so that $z_H(u) - z_G(u) = ab/(b+1) - c/(b+1)$. Thus,

$$\begin{aligned} L(G, \beta, W) - L(H, \beta, W) - Lin_{W, \beta}(G-H) &= L_1 - L_2 - L_3 + L_4 \\ &:= \iint W \left[\frac{ab}{b+1} - \frac{c}{b+1} \right] d(G-H) - \iint \int_0^t \left[\frac{ab}{b+1} - \frac{c}{b+1} \right] du z' d(G-H) \beta, \end{aligned}$$

where after distributing the inner brackets L_1 , through L_4 are each of the integral terms above. By assumption A2), we can choose large enough values M_0, M_1, \dots, M_p , so that $\tilde{K} := \prod_{i=0}^p [0, M_i]$ is bounded and it includes the support of any function in \mathcal{W}^* . Observe further that $[0, M_0]$ includes K .

Take $u \in [0, M]$ and $W \in \mathcal{W}$. Then,

$$\begin{aligned} \iint W(u, z) z I(t \geq u) dH(t, z) &= \iint_{\tilde{K}} W_0(u) W^*(u, z) z I(t \geq u) dH(t, z) \\ &= W_0(u) \int_0^{M_p} \dots \int_0^{M_1} W^*(u, z) z \left[\int_0^M I(t \geq u) \frac{\partial}{\partial t} \left(\frac{\partial^p}{\partial z} H(t, z) \right) dt \right] dz, \end{aligned}$$

where $\frac{\partial^p}{\partial z} = \frac{\partial^p}{\partial z_1 \dots \partial z_p}$ and $dz = dz_1 \dots dz_p$. The integral within the argument

$$\begin{aligned} \int_0^{M_0} I(t \geq u) \frac{\partial}{\partial t} \left(\frac{\partial^p}{\partial z} H(t, z) \right) dt &= I(M_0 > u) \int_u^{M_0} \frac{\partial}{\partial t} \left(\frac{\partial^p}{\partial z} H(t, z) \right) dt \\ &= I(M_0 > u) \left[\frac{\partial^p}{\partial z} H(M_0, z) - \frac{\partial^p}{\partial z} H(u, z) \right] \\ &= \Delta_{[u, M_0]} f(t) := I(M_0 > u) [f(M_0) - f(u)], \end{aligned}$$

where in order to simplify notation, we introduced the operator $\Delta_{[u, M_0]} f(t) := I(M_0 > u) [f(M_0) - f(u)]$. Denote also the set $\prod_{i=1}^p [0, M_i]$ by K_z and the integrals $\int_0^{M_p} \dots \int_0^{M_1}$ by \int_{K_z} , so that

$$\int_{K_z} W^*(u, z) z \int_0^{M_0} I(t \geq u) \frac{\partial}{\partial t} \frac{\partial^p}{\partial z} H(t, z) dt dz = \Delta_{[u, M_0]} \int_{K_z} W^*(u, z) z \frac{\partial^p}{\partial z} H(t, z) dz.$$

Since we chose the M_i large enough, $W^*(\dots, M_i, \dots) = W^*(\dots, 0, \dots) = 0$. By iterative integration by parts,

$$\Delta_{[u, M_0]} \int_{K_z} W^*(u, z) z \frac{\partial^p}{\partial z} H(t, z) dz = (-1)^p \Delta_{[u, M_0]} \int_{K_z} H(t, z) \frac{\partial^p}{\partial z} (W^*(u, z) z) dz.$$

In consequence,

$$\begin{aligned} a(u) &= \frac{W_0(u) (-1)^p \Delta_{[u, M_0]} \int_{K_z} H(t, z) \frac{\partial^p}{\partial z} (W^*(u, z) z) dz}{W_0(u) \iint W^*(u, z) I(t \geq u) dH(t, z)} \\ &= \frac{(-1)^p \Delta_{[u, M_0]} \int_{K_z} H(t, z) \frac{\partial^p}{\partial z} (W^*(u, z) z) dz}{\iint W^*(u, z) I(t \geq u) dH(t, z)}. \end{aligned}$$

Hence for all $u \in K$,

$$\|a(u)\|_2^2 \leq \frac{\|H\|_\infty^2 \left\| \int_{K_z} \frac{\partial^p}{\partial z} (W^*(u, z) z) dz \right\|_2^2}{\iint W^*(u, z) I(t \geq u) dH(t, z)},$$

which is bounded because of A1) and A2). *i.e.* for some constant, $\|a(u)\|_2 \leq c_a$. Similarly by integration by parts,

$$\begin{aligned} &\iint_{\bar{K}} W(u, z) I(t \geq u) d(G - H)(t, z) \\ &= W_0(u) (-1)^p \Delta_{[u, M_0]} \int_{K_z} (G - H)(t, z) \frac{\partial^p}{\partial z} W^*(u, z) dz. \end{aligned}$$

Thus

$$b(u) = \frac{(-1)^p \Delta_{[u, M_0]} \int_{K_z} (G - H)(t, z) \frac{\partial^p}{\partial z} W^*(u, z) dz}{\iint W^*(u, z) I(t \geq u) dH(t, z)},$$

so that $|b(u)| \leq \varepsilon^{-1} \|G - H\|_\infty \int_{K_z} \left| \left(\frac{\partial^p}{\partial z} \right) W^*(u, z) \right| dz$. By assumption A2) this integral is bounded. *i.e.* for all $u \in K$, $|b(u)| \leq c_b \|G - H\|_\infty$ for some finite constant c_b . A similar application of integration by parts and the assumptions gives that for all $u \in K$, $\|c(u)\|_2 \leq c_c \|G - H\|_\infty$ for some finite constant c_c . Lastly, we apply the same methodology to the integrals

$$\begin{aligned} S_1 &:= \iint W_0(t) W^*(t, z) d(G - H)(t, z) \\ S_2 &:= \iint W_0(t) W^*(t, z) \int I(t \geq u) du d(G - H)(t, z) \\ &= \iint W_0(t) W^*(t, z) t d(G - H)(t, z) \end{aligned}$$

to claim that for some finite constants s_1 and s_2 , $|S_1| \leq s_1 \|G - H\|_\infty$ and $|S_2| \leq s_2 \|G - H\|_\infty$. Also for $z \in K_z = \prod_{i=1}^p [0, M_i]$, which closed and bounded, we have that $\|\beta'z\|_\infty < \infty$. With the above bounds, we now focus on the terms L_1 through L_4

$$\begin{aligned} \|L_1\|_2 &= \left\| \iint W_0(t) W^*(t, z) a(t) \frac{b(t)}{b(t) + 1} d(G - H)(t, z) \right\|_2 \\ &\leq \|a(t)\|_2 \|b(t)\| |S_1| \sqrt{p} \leq c_a c_b \|G - H\|_\infty s_1 \|G - H\|_\infty \sqrt{p} = o\left(\|G - H\|_\infty^2\right). \end{aligned}$$

Similar calculations hold for the other L_i 's. Detailed calculations can be found in Juieta Ferrario's Ph.D. dissertation [15]. Finally, joining all bounds together we can assert that

$$\sup_{\substack{\beta \in B \\ W \in \mathcal{W}}} \|L(G, \beta, W) - L(H, \beta, W) - \text{Lin}_{W, \beta}(G - H)\|_2 = o(\|G - H\|_\infty^2).$$

For the last assertion of the Lemma, substitute H by $(G - H)$ in the functions $b(u)$ and $c(u)$ to obtain that $b(u) = 1$ and $c(u) = a(u)$. With that, $z_H(u) - z_G(u) = 0$, which cancels out all but the first two terms in $\text{Lin}_{W, \beta_0}(H)$, which entails by Fisher-consistency that $\text{Lin}_{W, \beta_0}(H) = L(H, \beta_0, W) = 0$.

Proof of Lemma 2. For any fixed $W \in \mathcal{W}$ and $\beta \in B$,

$$\frac{\partial L(H, \beta, W)}{\partial \beta} = - \left[\iint W(t, z) \left(\int_0^t [z - z_H(u)] du \right) z' dH(t, z) \right] = \mathcal{L}(H, W),$$

which is independent of β . With the above we express

$$L(H, \beta, W) = \iint W(t, z) [z - z_H(t)] dH(t, z, \Delta = 1) + \mathcal{L}(H, W) \beta,$$

$$L(H, \beta_0, W) = \iint W(t, z) [z - z_H(t)] dH(t, z, \Delta = 1) + \mathcal{L}(H, W) \beta_0.$$

So subtracting, $L(H, \beta, W) - L(H, \beta_0, W) = \mathcal{L}(H, W)(\beta - \beta_0)$.

Proof of Lemma 3. Express

$$\begin{aligned} \text{Lin}_{W, \beta}(G - H) &= \iint W[z - a] d(G - H) - W \left(\int_0^t [z - a] du \right) z' d(G - H) \beta \\ &\quad + \iint W \left[\frac{ab}{b+1} - \frac{c}{b+1} \right] dH - \iint W \left(\int_0^t \left[\frac{ab}{b+1} - \frac{c}{b+1} \right] du \right) z' dH \beta \\ &=: R_1 - R_2 - R_3 + R_4 + \tilde{L}_1 - \tilde{L}_2 - \tilde{L}_3 + \tilde{L}_4, \end{aligned}$$

where $R_1 := \iint W_0(t) W^*(t, z) z d(G - H)(t, z)$ and the other integrals are defined similarly. Following the same arguments as in Lemma 1, relying on integration by parts and Assumptions A1)-A2), we see that all the terms are $o(\|G - H\|_\infty)$, which finishes the proof. Detailed calculations are shown in [15].

Proof of Theorem 1. By Lemmas 1 and 3, for all $W \in \mathcal{W}$ and $\beta \in B$, there exists some constant M_0 so that

$$\|L(G, \beta, W) - L(H, \beta, W)\|_2 \leq M_0 \|G - H\|_\infty. \quad (16)$$

Also by Lemma 2, $L(H, \beta, W) - L(H, \beta_0, W) - \mathcal{L}(H, W)(\beta - \beta_0) = 0$, so that adding to (16) we get $\|L(G, \beta, W) - \mathcal{L}(H, W)(\beta - \beta_0)\|_2 \leq M_0 \|G - H\|_\infty$, for $L(H, \beta_0, W) = 0$. Since by Assumption A3),

$a := |\det \{\mathcal{L}(H, W)\}|$ is finite and positive, this ensures that

$$\|\mathcal{L}^{-1}(H, W) L(G, \beta, W) - (\beta - \beta_0)\|_2 \leq M_0 \|G - H\|_\infty a^{-1/p}. \quad (17)$$

Take now $M := M_0 a^{-1/p} > 0$ and define a continuous function $g : B \mapsto B$ by

$$g(\beta) := \beta - \mathcal{L}^{-1}(H, W) L(G, \beta, W),$$

for some fixed $H \in \mathcal{A}$, $G \in \mathcal{B}(H, \varepsilon)$ and $W \in \mathcal{W}$. Thus by (17) we have that

$$\|\beta_0 - g(\beta)\|_2 \leq M \|G - H\|_\infty. \quad (18)$$

So for $\|G - H\|_\infty$ sufficiently small, the ball $\{\beta : \|\beta - \beta_0\|_2 \leq M \|G - H\|_\infty\}$ is a subset of B . Also, by Equation (18), if some $\beta \in B$, its image $g(\beta) \in B$ too. By Brouwer's fixed point theorem, there exists $\beta^* \in B$ for which $g(\beta^*) = \beta^*$, i.e.

$$\beta^* - \mathcal{L}^{-1}(H, W) L(G, \beta^*, W) = \beta^*,$$

which implies that $L(G, \beta^*, W) = 0$.

Proof of Theorem 2. Since $L(H, \beta_0, W) = 0$,

$$L(G, \beta, W) = L(G, \beta, W) - L(H, \beta, W) + L(H, \beta, W) - L(H, \beta_0, W). \quad (19)$$

Also by Lemmas 1 and 2 respectively, for all $W \in \mathcal{W}$ and $\beta \in B$,

$$L(G, \beta, W) - L(H, \beta, W) = \text{Lin}_{W, \beta_0}(G - H) + o(\|G - H\|_\infty),$$

$$L(H, \beta, W) - L(H, \beta_0, W) = \mathcal{L}(H, W)(\beta - \beta_0).$$

Adding the above Equations in (19) we get

$$L(G, \beta, W) = \text{Lin}_{W, \beta_0}(G - H) + \mathcal{L}(H, W)(\beta - \beta_0) + o(\|G - H\|_\infty).$$

Note that by Theorem 1, for $\|G - H\|_\infty$ sufficiently small, there exists $\beta(G, W)$ so that $\beta(G, W) - \beta_0 = o(\|G - H\|_\infty)$ and $L(G, \beta(G, W), W) = 0$. *i.e.*

$$\mathcal{L}(H, W)\beta(G, W) - \beta_0 + \text{Lin}_{W, \beta_0}(G - H) = o(\|G - H\|_\infty). \quad (20)$$

Now since by Assumption A3) $\det \mathcal{L}(H, W)$ is nonzero and bounded on \mathcal{W} , premultiplying by the inverse we get

$$\beta(G, W) - \beta_0 + \mathcal{L}(H, W)^{-1} \text{Lin}_{W, \beta_0}(G - H) = o(\|G - H\|_\infty). \quad (21)$$

Proof of Theorem 3. Decompose $\|\mathbb{H}_n - H\|_\infty \leq \|\mathbb{H}_n - G\|_\infty + \|G - H\|_\infty$. By assumption $\|G - H\|_\infty = o(n^{-1/2})$ and by Glivenko-Cantelli's Theorem $\|\mathbb{H}_n - G\|_\infty = o_p(n^{-1/2})$. Then by Theorem 2,

$$\beta(\mathbb{H}_n, W) - \beta_0 + \mathcal{L}(H, W)^{-1} \text{Lin}_{W, \beta_0}(\mathbb{H}_n - H) = o_p(n^{-1/2}), \quad (22)$$

$$n^{1/2} \left[\beta(\mathbb{H}_n, W) - \beta_0 + \mathcal{L}(H, W)^{-1} \text{Lin}_{W, \beta_0}(\mathbb{H}_n - H) \right] = o_p(1). \quad (23)$$

Since $\text{Lin}_{W, \beta_0}(\mathbb{H}_n - H)$ is linear and $\text{Lin}_{W, \beta_0}(H - H) = 0$, the conclusion follows by the classical Central Limit's Theorem.