Harvard University

Harvard University Biostatistics Working Paper Series

Year 2006 *Paper* 48

Predicting Future Responses Based on Possibly Misspecified Working Models

Tianxi Cai[*] Lu Tian[†]

Scott D. Solomon[‡] L.J. Wei[**]

[*]Harvard University, tcai@hsph.harvard.edu

[†]Northwestern University, lutian@northwestern.edu

[‡]Brigham & Women's Hospital, ssolomon@rics.bwh.harvard.edu

[**]Harvard University, wei@sdac.harvard.edu

PREDICTING FUTURE RESPONSES BASED ON POSSIBLY MISSPECIFIED WORKING MODELS

Tianxi Cai
Department of Biostatistics, Harvard University
677 Huntington Ave.
Boston, Massachusetts 02115, U.S.A.
tcai@hsph.harvard.edu

Lu Tian
Department of Preventive Medicine, Northwestern University
680 N. Lake Shore Drive
Chicago, Illinois 60611, U.S.A.
lutian@northwestern.edu

Scott D. Solomon,
Cardiovascular Division, Brigham & Women's Hospital
75 Francis St.
Boston, MA 02115, U.S.A.
ssolomon@rics.bwh.harvard.edu

and L.J. Wei
Department of Biostatistics, Harvard University
677 Huntington Ave.
Boston, Massachusetts 02115, U.S.A.
wei@sdac.harvard.edu

SUMMARY

Under a general regression setting, we propose an optimal *unconditional* prediction procedure for future responses. The resulting prediction intervals or regions have a desirable *average* coverage level over a set of covariate vectors of interest. When the working model is not correctly specified, the traditional conditional prediction method is generally invalid. On the other hand, one can empirically calibrate the above unconditional procedure and also obtain its cross-validated counterpart. Various large and small sample properties of these unconditional methods are examined analytically and numerically. We find that the $\mathcal{K}$-fold cross validated procedure performs exceptionally well even for cases with rather small sample sizes. The new proposals are illustrated with two real examples, one with a continuous response and the other with a binary outcome.

Keywords: Heterogeneous regression; $\mathcal{K}$-fold cross validation; Misspecified regression model; Optimal prediction region; Prediction error rate.

1. INTRODUCTION

One of the main goals of regression analysis is to predict future responses based on vectors of observable "baseline" covariates (Patel, 1989; Geisser, 1993; Preston, 2000). The conventional frequentist's prediction interval or region guarantees a certain coverage level under the setting that one would repeatedly draw future subjects with the *same fixed* covariate vector of interest (Stine, 1985; Carroll & Ruppert, 1991; Schmoyer, 1992; Olive, 2006). However, this conditional coverage level requirement is rather stringent and may not be practically relevant. In practice, a prediction interval procedure will be used for predicting future responses repeatedly for various distinct sets of covariates. Therefore, it is appealing to consider prediction regions which have the desirable *average* coverage level with respect to the covariate vector from a population of interest. Similar arguments for utilizing such an *averaging* concept to evaluate a general statistical method, which is expected to be used repeatedly under different settings, have been made by Neyman (1977), Rubin (1984), Bayarri & Berger (2004) and Uno, Tian & Wei (2005).

In the first part of this article, we assume that the working regression model is correctly specified and show how to construct an "optimal" prediction interval procedure among all the aforementioned unconditional prediction methods with a pre-specified coverage level. In the second part of the paper, we consider the situation that the working model may be misspecified. For this case, generally the traditional conditional prediction intervals are not valid. On the other hand, one can empirically calibrate the above unconditional optimal prediction procedure and also obtain its $\mathcal{K}$-fold cross-validated counterpart. We demonstrate that the true average coverage probability of the resulting calibrated prediction regions converges to the nominal level and that the sampling distribution of the true coverage level can be approximated by a simple normal distribution. This approximation can be readily used to assess the reliability of a prediction procedure. An extensive numerical study is also conducted to examine the finite sample properties of the new methods. We find that the empirically calibrated, $\mathcal{K}$-fold cross validated procedure performs exceptionally well even for cases with rather small sample sizes. All the proposals are illustrated with two real examples, one with a continuous response variable and another with a binary outcome.

Note that when the fitted regression model is misspeficied, robust inference procedures for regression parameters, especially with respect to hypothesis testing, have been studied, for example, by Gail, Wieand & Piantadosi (1984), Lagakos & Schoenfeld (1984), Struthers & Kalbfleisch (1986) and Lin & Wei (1989).

2. OPTIMAL PREDICTION INTERVALS OR REGIONS WHEN THE WORKING MODEL IS CORRECTLY SPECIFIED

First we consider the case that the response variable Y is absolutely continuous. Let Z be its p-dimensional bounded covariate vector whose first component is one. Also, let Θ be a vector of unknown parameters, whose dimension is either infinite or finite. Assume that there exists $\Theta = \Theta_0$ such that for $Z = z$, the *conditional* distribution of Y can be generated via a random variable $Y_{\Theta_0}(z)$. For example, one may let

$$h\{Y_\Theta(z)\} = g(\beta' z) + \sigma(\gamma' z)\epsilon, \tag{2.1}$$

where β and γ are unknown parameter vectors, $h(\cdot)$, $g(\cdot)$, and $\sigma(\cdot)$ are pre-specified strictly monotone functions, $\sigma(\cdot) > 0$, and ϵ is a random error term which is free of z with zero mean and unit variance. This is a typical heterogeneous regression model, which relates a continuous response to its covariates (Carrol & Ruppert, 1988). If the distribution form of ϵ is completely unspecified, Θ consists of β, γ and the distribution or density function of ϵ.

Suppose that we are interested in predicting the response Y^0 of a future subject with covariate vector $Z^0 = z^0$. Moreover, suppose that the conditional distribution of Y^0 given z^0 is the same as that of $Y_{\Theta_0}(z^0)$. Then a *theoretical* prediction interval or region $J_\eta(z^0)$ for Y^0 with coverage level $0 < \eta < 1$ is a set of possible values for Y^0 such that

$$\text{pr}(Y^0 \in J_\eta(z^0) | Z^0 = z^0) = \text{pr}(Y_{\Theta_0}(z^0) \in J_\eta(z^0)) = \eta. \tag{2.2}$$

Like the standard confidence or credible regions, there are many choices of $J_\eta(z^0)$. On the other hand, it is not difficult to obtain a prediction region which has the smallest size among all $J_\eta(z^0)$. To this end, let $f(y; z^0)$ be the continuous density function of $Y_{\Theta_0}(z^0)$, and

$$I_\eta(z^0) = \{y : f(y; z^0) \geq c_\eta(z^0)\}, \tag{2.3}$$

where $c_\eta(z^0)$ is chosen such that $I_\eta(z^0)$ satisfies (2.2). If $\mathrm{pr}(f(Y^0; Z^0) = s \mid Z^0 = z^0) = 0$ for any $s > 0$, then such $c_\eta(z^0)$ uniquely exists. It follows from the argument for the optimality property of the highest posterior density region in the Bayesian literature (Box & Tiao, 1973, pp.123-124) that $I_\eta(z^0)$ is the optimal one in the sense that $\|I_\eta(z^0)\| \leq \|J_\eta(z^0)\|$, where $\|\mathcal{A}\|$ denotes the length or size of the set $\mathcal{A}$.

Now, to obtain *empirical* prediction regions for Y^0, assume that the data $\{(Y_i, Z_i),\ i = 1, \cdots, n\}$ are n independent copies of (Y, Z) and let $\hat{\Theta}$ be a "consistent" estimator of Θ_0. The mean or mode of $Y_{\hat{\Theta}}(z^0)$ is a reasonable point estimate for Y^0. An η-level empirical region $\hat{J}_\eta(z^0)$ corresponding to (2.2) is defined as

$$\mathrm{pr}(Y_{\hat{\Theta}}(z^0) \in \hat{J}_\eta(z^0) |\ \text{Data}) = \eta, \tag{2.4}$$

where the probability is with respect to $Y_{\hat{\Theta}}(z^0)$ given the data. Under certain regularity conditions, the true coverage level of $\hat{J}_\eta(z^0)$ converges to η. That is, as a function of the data, $\mathrm{pr}(Y^0 \in \hat{J}_\eta(z^0) |\ Z^0 = z^0, \text{Data})$, converges to η, in probability, as $n \to \infty$.

Now, let $\hat{f}(\cdot; z^0)$ be the density function of $Y_{\hat{\Theta}}(z^0)$. Then, the empirical counterpart for (2.3) is

$$\hat{I}_\eta(z^0) = \{y : \hat{f}(y; z^0) \geq \hat{c}_\eta(z^0)\}, \tag{2.5}$$

where $\hat{c}_\eta(z^0)$ is chosen such that $\hat{I}_\eta(z^0)$ satisfies (2.4). Assume that $\hat{f}(y; z^0)$ converges to $f(y; z^0)$, uniformly in y, as $n \to \infty$, then it is straightforward to show that $\|\hat{I}_\eta(z^0)\|$ converges to $\|I_\eta(z^0)\|$ in probability. Therefore, in the limit, $\hat{I}_\eta(z^0)$ is expected to the smallest region among all $\hat{J}_\eta(z^0)$. When Model (2.1) is the standard linear regression with identically distributed error terms and

$f(\cdot; z^0)$ is unimodal, we expect that in the limit, the set $\hat{I}_\eta(z^0)$ corresponds to the optimal interval studied by Olive (2006) derived via the empirical distribution function of the residuals. It is important to note that when the fitted model is misspecified, asymptotically the coverage probability of $\hat{I}_\eta(z^0)$ can be quite different from its nominal coverage level η.

Note that Condition (2.4) imposed on the prediction intervals $\hat{J}_\eta(\cdot)$ is rather stringent and may not be practically relevant. It guarantees the validity of $\hat{J}_\eta(z^0)$ under the setting that one would repeatedly draw future subjects with the same *fixed* z^0. In practice, a prediction interval procedure will be used for predicting Y^0 with many different values of Z^0, say, in a set $\mathcal{D}$. Therefore, it is appealing to consider prediction region $\hat{L}_\eta(\cdot)$ such that

$$\text{pr}(Y_{\hat{\Theta}}(Z^0) \in \hat{L}_\eta(Z^0) \mid \text{Data}) = \int_{\mathcal{D}} \text{pr}(Y_{\hat{\Theta}}(z) \in \hat{L}_\eta(z) \mid \text{Data}) dH(z) = \eta, \tag{2.6}$$

where $H(\cdot)$ is the distribution function of $Z^0 \in \mathcal{D}$. In Appendix A, we show that the true average coverage level of $\hat{L}_\eta(\cdot)$, $\text{pr}(Y^0 \in \hat{L}_\eta(Z^0) \mid \text{Data})$, converges to η, in probability, as $n \to \infty$, where the probability is with respect to the joint distribution of Y^0 and Z^0.

Note that when the fitted model is not correctly specified, the unconditional prediction procedure $\hat{L}_\eta(\cdot)$ can be calibrated empirically so that the average coverage level of the resulting regions is about η. On the other hand, it is rather difficult, if not impossible, to do so for its conditional counterpart $\hat{J}_\eta(\cdot)$. More details are given in Section 3.

Any conditional prediction set $\hat{J}_\eta(\cdot)$ in (2.4) automatically satisfies (2.6). On the other hand, the class of prediction sets $\hat{L}_\eta(\cdot)$ is much larger than that of $\hat{J}_\eta(\cdot)$. The question is how to identify an "optimal" procedure, for example, which produces prediction regions with the smallest average length or size over $Z^0 \in \mathcal{D}$ among all $\hat{L}_\eta(\cdot)$. This seems to be a rather complex optimization problem. It turns out that such an optimal region can be constructed quite easily. Specifically, let

$$\hat{K}_\eta(z) = \{y : \hat{f}(y; z) \geq \hat{c}_\eta\}. \tag{2.7}$$

Here, $\hat{c}_\eta$ is $free$ of z and is chosen such that (2.6) is satisfied, that is,

$$\int_{\mathcal{D}} \int_{-\infty}^{\infty} \mathcal{I}\{\hat{f}(y;z) \geq \hat{c}_\eta\} \hat{f}(y;z) dy dH(z) = \eta,$$

where $\mathcal{I}(\cdot)$ is the indicator function. In Appendix A, we show that under some regularity conditions, if $\sup_{y,z} |\hat{f}(y;z) - f(y;z)| \to 0$ in probability, as $n \to \infty$, the limit of $\mathrm{E}(\|\hat{K}_\eta(Z^0)\|)$ is no greater than that of $\mathrm{E}(\|\hat{L}_\eta(Z^0)\|)$, where the expectation is with respect to $Z^0 \in D$.

Now, let us use an example with a relatively small sample size to illustrate the proposed procedures. The data set of this example consists of 54 patient records (Neter, Wasserman & Kutner, 1985, p.419). Each record has the patient's survival time Y after a liver surgery and the corresponding four pre-operational biomarker values: blood clotting score (BCScore), prognostic index (PIndex), enzyme function test score (EScore) and liver function test score (LScore). Here, Z is a 5 × 1 vector. The goal is to establish a prediction model for the patient's survival after the surgery via these four pre-operational covariates. We used Model (2.1) to fit these data with $h(x) = \log(x)$ and $g(x) = x$, $\sigma(x) = \exp(x)$, and an unspecified distribution of the error term ϵ. Here, Θ_0 consists of the true values β_0 and γ_0 of β and γ, and the true density function of ϵ. The estimator $\hat{\beta}$ for β_0, is obtained via the following simple estimating function

$$\sum_{i=1}^{n} Z_i\{\log(Y_i) - g(\beta' Z_i)\}. \tag{2.8}$$

We then estimate γ_0 by $\hat{\gamma}$ with the estimating function

$$\sum_{i=1}^{n} Z_i \left[\{\log(Y_i) - g(\hat{\beta}' Z_i)\}^2 - \sigma^2(\gamma' Z_i)\right]. \tag{2.9}$$

The resulting estimates $\hat{\beta}$ and $\hat{\gamma}$ and the corresponding estimated standard errors are given in Table 1. Note that more efficient estimation procedures for Model (2.1) may be used for estimating β_0 and γ_0. On the other hand, when the fitted model is not correctly specified, the estimators obtained via (2.8) and (2.9) are "well-behaved". More details are given in the next section.

Lastly, for the present example, we estimate the density function of the error term ϵ via the following univariate kernel function estimate $\hat{f}_0(\cdot)$ with the standardized residuals, $\{\hat{e}_i, i = 1, \cdots, n\}$, where

$$\hat{f}_0(x) = n^{-1} \sum_{i=1}^{n} \phi_\tau(\hat{e}_i - x),$$

$\hat{e}_i = \{h(Y_i) - g(\hat{\beta}' Z_i)\}/\sigma(\hat{\gamma}' Z_i)$, $\phi_\tau(x) = \tau^{-1}\phi(x/\tau)$, $\phi(\cdot)$ is the standard normal density function and τ is the smooth parameter. Like any nonparametric function estimation problem, the proper choice of the smooth parameter τ is not obvious. In practice, one may use the simple rule-of-thumb proposed by Scott (1992) for choosing the bandwidth, that is, $\tau = 1.06n^{-1/5} \times \min(1, 1.34\text{IQR})$, where IQR is the interquartile range of $\{\hat{e}_i, i = 1, \cdots, n\}$. Alternatively, one may select an optimal bandwidth based on cross-validation methods by minimizing the mean square error of the resulting density estimator (Rudemo, 1982; Bowman, 1984). For all numerical studies discussed in the paper, we use the cross-validation method to select the bandwidth τ.

With the above estimates, conditional on $Z = z$, the prediction density function $\hat{f}(y; z)$ of $Y_{\hat{\Theta}}(z)$ is

$$\hat{f}(y; z) = \frac{\dot{h}(y)}{\sigma(\hat{\gamma}' z)} \hat{f}_0 \left(\frac{h(y) - g(\hat{\beta}' z)}{\sigma(\hat{\gamma}' z)} \right) = \frac{1}{y \exp(\hat{\gamma}' z)} \hat{f}_0 \left(\frac{\log(y) - \hat{\beta}' z}{\exp(\hat{\gamma}' z)} \right), \tag{2.10}$$

where $\dot{h}(y)$ is the derivative of $h(y)$. In Appendix B, we show under a rather general setting that $\hat{\Theta}$ is consistent under Model (2.1). To illustrate how to construct prediction intervals $\hat{I}_\eta(\cdot)$ and $\hat{K}_\eta(\cdot)$, in Figure 1, we plot the prediction density functions $\hat{f}(\cdot; z^0)$ with two distinct sets of z^0. The density function on the right hand side, labeled by (a), is relatively flat and skewed. The 0.8 conditional prediction interval $\hat{I}_{0.8}(z^0)$ is (328 days, 424 days), a quite large interval. Now, if we let the distribution $H(\cdot)$ of z^0 in (2.6) be the empirical distribution of $\{Z_i, i = 1, \cdots, n\}$, the corresponding 0.8 unconditional region $\hat{K}_{0.8}(z^0) = (346, 403)$, which is a relatively tight interval. Here, $\hat{c}_{0.8} = 0.0063$ and $\hat{c}_{0.8}(z^0) = 0.0020$. The prediction density function labeled by (b) is narrow and peaky. The corresponding sets $\hat{I}_{0.8}(z^0)$ and $\hat{K}_{0.8}(z^0)$ are $(110, 136) \cup (144, 145)$ and $(110, 137) \cup (140, 150)$, respectively. For this case, the unconditional region is slightly larger than its conditional counterpart, but it is still tight enough for making practically meaningful predictions.

For a global comparison between these two interval procedures with this example, in Figure 2, we provide a scatter diagram with 54 dots, the x and y of each dot denote the sizes of $\hat{I}_{0.8}(z^0)$ and $\hat{K}_{0.8}(z^0)$ for a study subject with its observed preoperational covariate vector z^0. The average sizes of $\hat{I}_{0.8}$ and $\hat{K}_{0.8}$ over these 54 cases are 43 and 38 days, respectively.

Note that one can also fit the data with a standard normal error ϵ in Model (2.1). The resulting conditional prediction interval $\hat{I}_{0.8}(z^0)$ for Case (a) presented in Figure 1 is (326, 430), and for the second case, it is (110, 138). Their unconditional counterparts are (359, 391) and (106, 142), respectively. The average lengths over 54 cases are 45 days for the conditional intervals and 40 days for the unconditional intervals.

Now, let us consider the case that the response Y is discrete. Like the continuous case, we show in Appendix A that asymptotically the unconditional set $\hat{K}_\eta(\cdot)$ has the smallest average size with respect to Z among all $\hat{L}_\eta(\cdot)$ provided that η is an attainable prediction level of these sets. Here the size is determined by the counting measure. Note that when there is at least one continuous covariate in Z, for any given $0 < \eta < 1$, in general one can obtain $\hat{K}_\eta(\cdot)$, however, its conditional counterpart $\hat{I}_\eta(\cdot)$ may not exist.

Let us use an example with a binary outcome to illustrate how to construct the prediction set $\hat{K}_\eta(\cdot)$. The data set of the example is from a study called "HEART", the Healing and Early Afterload Reducing Therapy Trial (Pfeffer et al., 1997; Manes et al., 2003), which is a randomized double-blind study of the hemodynamic effects of early versus delayed administration of ramipril after myocardial infarction. Although there were no significant differences with respect to the patient's mortality or morbidity among three treatment groups in the trial, it is interesting to use the data to establish a prediction model for early identification of high risk patients for proper medical interventions or for planning future studies. Here, the response is binary, which is one if the patient died or had a heart failure by or at a year after randomization. The covariates are the patient's 14 day ejection fraction (EFrac), average ST-segment evaluation at day 7 (AveST), and maximum ST-segment elevation at day 7 (MaxST). There are 274 study patients who have complete information about these variables. Here, we assume that $Y_\Theta(z)$ is a binary variable with the failure probability $\text{pr}(Y = 1|\ Z = z) = g(\beta' z)$. The parameter vector $\Theta = \beta$. For this present example, we let $g(x) = \{1+\exp(-x)\}^{-1}$, the standard logistic link function. A consistent estimator $\hat{\beta}$ is obtained via the estimating function (2.8) with identity function $h(\cdot)$. It follows that given $Z = z$, the prediction density function $\hat{f}(y; z)$ for the binary variable $Y_{\hat{\Theta}}(z)$ is

$$yg(\hat{\beta}' z) + (1 - y)(1 - g(\hat{\beta}' z)). \tag{2.11}$$

The η-level prediction set $\hat{K}_\eta(\cdot)$ in (2.7) can then be constructed accordingly via (2.11).

It is interesting and important to examine the connection between $\hat{K}_\eta(\cdot)$ and the classical binary classification rule. For a future patient with covariate vector z^0, a conventional classification rule predicts $Y^0 = 1$, if $Y_{\hat{\Theta}}(z^0) \geq \xi$, otherwise, $Y^0 = 0$, where ξ is chosen between 0 and 1 to satisfy certain criteria. For this classification rule, the corresponding prediction set $\hat{L}_\eta(z^0)$ in (2.6) consists of a single element, either 0 or 1, where η corresponds to the conventional correct classification probability. Note that this rule may not produce the best prediction set $\hat{K}_\eta(\cdot)$ defined in (2.7) due to the fact that a general η prediction set has four possibilities, the empty set, $\{0\}, \{1\}$ and $\{0, 1\}$.

A commonly used classification rule is the one with $\xi = 0.5$. For the HEART trial, the corresponding prediction set attains $\eta = 0.82$ and coincides with $\hat{K}_{0.82}(\cdot)$. Thus, when the fitted model is correctly specified, this specific rule gives the best prediction set among all 0.82-level prediction sets. On the other hand, if one would like to have a higher prediction level, say, $\eta = 0.9$, then the above rule is no longer valid.

For illustration, suppose that we are interested in predicting future responses for two distinct sets of covariate vectors. The $z^0 = (1, \text{EFrac}, \text{AveST}, \text{MaxST})'$ of the first case is $(1, 68, 0.52, 1.16)'$ and for the second case is $(1, 35, 0.5, 1.26)'$. For the first case, the estimate of the failure probability

$g(\hat{\beta}'z^0)$ is 0.04, a very small value. The $\hat{K}_{0.82}(z^0) = \{0\}$, indicating that very likely the future patient is free of the event. For the second case, the probability of failure is 0.51, and $\hat{K}_{0.82}(z^0) = \{1\}$. On the other hand, if we choose $\eta = 0.9$, a relatively high coverage level, the corresponding $\hat{c}_{0.9} = 0.27$ instead of 0.5. Now, for the first case, $\hat{K}_{0.9}(z^0)$ is still $\{0\}$. However, for the second case, $\hat{K}_{0.9}(z^0) = \{0, 1\}$. This suggests that although we cannot make a good prediction based on three "baseline" covariates, we will provide extra, maybe quite costly medical interventions to this type of subjects for preventing them from early heart failure or death. Therefore, a high η value is associated with high medical cost.

Now, suppose that the resource is limited and we are willing to consider a prediction rule with a relatively low prediction level, say, $\eta = 0.7$. This results in $\hat{c}_{0.7} = 0.72$. Again, for the first case discussed above, $\hat{K}_{0.7}(z^0)$ is still $\{0\}$. However, for the second case, $\hat{K}_{0.7}(z^0)$ becomes an empty set, that is, we will not do anything for this type of subjects, but allocate resource to subjects with non-empty prediction sets. There are 53 empty prediction sets with the HEART data set. For those patients with non-empty prediction sets, the correct classification rate is 0.87, which is higher than the correct classification rate for the standard binary decision rule with the cutoff value of 0.5.

3. PREDICTION INTERVALS OR REGIONS WHEN THE WORKING MODEL MAY NOT BE CORRECTLY SPECIFIED

In practice, the conditional distribution of $Y_{\Theta_0}(z)$ is simply an approximation to the true conditional distribution of Y given $Z = z$. Therefore, even asymptotically, the coverage probability of a prediction set $\hat{L}_\eta(\cdot)$ defined by (2.6) can be markedly different from its nominal level η. On the other hand, if the distribution $H(\cdot)$ of Z^0 in (2.6) is from the same population of the observed Z's in the data, one may consider an empirically calibrated prediction set $\tilde{L}_\eta(\cdot)$ such that

$$n^{-1} \sum_{i=1}^{n} \mathcal{I}(Y_i \in \tilde{L}_\eta(Z_i)) = \eta. \tag{3.1}$$

Correspondingly, let $\tilde{K}_\eta(\cdot)$ be defined by (2.7), but with a cutoff point $\tilde{c}_\eta$ chosen to satisfy

$$n^{-1}\sum_{i=1}^{n}\mathcal{I}\{\hat{f}(Y_i, Z_i) \geq \tilde{c}_\eta\} = \eta. \tag{3.2}$$

Note that when at least one of the covariates is continuous, it is difficult, if not impossible, to calibrate empirically the conditional interval $\hat{I}_\eta(z^0)$ so that its coverage level is approximately η for given z^0.

To illustrate how to construct $\tilde{K}_\eta(\cdot)$, consider the working model (2.1) for a continuous response variable Y with a completely unspecified density function of ϵ. Now, even when the model is misspecified, it follows from the argument in Appendix A of Tian et al. (2006) that as $n \to \infty$, $\hat{\beta}$ and $\hat{\gamma}$ obtained via the estimation functions (2.6) and (2.7) still converge, in probability, to finite constants, say, β_0 and γ_0, respectively. Furthermore, the *working* prediction density function $\hat{f}(\cdot; z)$ is still (2.10).

In Appendix C, we show that, under the possibly misspecified model (2.1), the true coverage level, $\tilde{\eta} = \mathrm{pr}(Y^0 \in \tilde{K}_\eta(Z^0)|$ Data), converges to η in probability, as $n \to \infty$, without assuming that the distribution of $Y_{\Theta_0}(z^0)$ is the true distribution of Y^0 given $Z^0 = z^0$. Moreover, for large n, the distribution of $n^{1/2}(\tilde{\eta}-\eta)$ can be approximated well by a normal with mean 0 and variance $\eta(1-\eta)$. This rather simple approximation can be used for identifying possible values of the true coverage level of $\tilde{K}_\eta(\cdot)$. Note that all the above large sample properties can be justified for the case when Y is discrete.

When the sample size n is not large with respect to the dimension of Z, the expected value of $\tilde{\eta}$ for the prediction set $\tilde{K}_\eta(\cdot)$ in (3.2) can be markedly different from its nominal level η. This is analogous to the bias issue regarding the "apparent error" estimator for the prediction error (Efron, 1986). One common remedy to reduce such bias is to use the cross-validation procedure. Here, we propose to use the $\mathcal{K}-$fold cross-validation approach to obtain prediction sets. Specifically, we randomly split the data into $\mathcal{K}$ disjoint subsets of about equal size and let $\xi_i \in \{1, ..., \mathcal{K}\}$ denote the group label for the ith subject, that is, $\xi_i = k$ represents that the ith subject falls into the group k. For each $k \in \{1, ..., \mathcal{K}\}$, we use all observations not in group k to obtain the estimator $\hat{f}_{(-k)}(y, z)$ in (2.10), and use the observations in group k to calibrate the coverage level. Specifically, we obtain an η-level prediction set $\tilde{K}_\eta^{cv}(\cdot)$ defined by (2.7), but its cutoff point $\hat{c}_\eta = \tilde{c}_\eta^{cv}$ satisfies

$$n^{-1} \sum_{k=1}^{\mathcal{K}} \sum_{\xi_i = k} \mathcal{I}\left\{\hat{f}_{(-k)}(Y_i, Z_i) \geq \tilde{c}_\eta^{cv}\right\} = \eta.$$

Let $\tilde{\eta}^{cv}$ denote the corresponding true coverage level of $\tilde{K}_\eta^{cv}(\cdot)$. We show in Appendix D that as $n \to \infty$, $n^{1/2}(\tilde{\eta}^{cv} - \eta)$ converges in distribution to $N(0, \eta(1-\eta))$, the limiting distribution of $n^{1/2}(\tilde{\eta} - \eta)$.

Now, with the liver surgery data, for the future patients with z^0 for Case (a) in Figure 1, the 0.8 prediction interval $\tilde{K}_{0.8}(z^0)$ is (349, 400), which is similar to $\hat{K}_{0.8}(z^0)$ obtained under the assumption that the fitted model is true. With the 5-fold cross-validated procedure, $\tilde{K}_{0.8}^{cv}(z^0) = (343, 406)$. For patients with z^0 For Case (b), the interval $\tilde{K}_{0.8}(z^0)$ is $(110, 136) \cup (144, 146)$ and $\tilde{K}_{0.8}^{cv}(z^0)$ is (109, 151). Here, $\tilde{c}_{0.8} = 0.0076$ and $\tilde{c}_{0.8}^{cv} = 0.0056$. Moreover, the distributions of $\tilde{\eta}^{cv}$ and $\tilde{\eta}$ are approximately normal with mean 0 and standard error 0.054. For the present small study, with 95% probability with respect to the sampling variation, the true coverage levels of $\tilde{K}_{0.8}(\cdot)$ and $\tilde{K}_{0.8}^{cv}(\cdot)$ are between 0.69 and 0.91. Note that one may increase the nominal level η to obtain an "acceptable" lower bound for $\tilde{\eta}$ and $\tilde{\eta}^{cv}$. The average length of $\tilde{K}_{0.8}^{cv}$ over these 54 cases is 42 days which is slightly shorter than that of $\hat{I}_{0.8}$.

One can also obtain empirically calibrated $\tilde{K}_\eta(\cdot)$ and $\tilde{K}_\eta^{cv}(\cdot)$ by fitting the data via Model (2.1) with the standard normal error. For the first case in Figure 1, the resulting $\tilde{K}_{0.8}(z^0) = (369, 379)$ and $\tilde{K}_{0.8}^{cv}(z^0) = (345, 405)$. For Case (b), the corresponding intervals are $(107, 142)$ and $(105, 144)$, respectively.

Now, for the case with a binary response Y, we use the logistic regression working model with failure probability $\text{pr}(Y = 1 \mid Z = z) = g(\beta' z)$ and the estimating function (2.8) with $h(\cdot)$ being the identity function. With the data from the HEART study discussed in Section 2, $\tilde{c}_{0.7} = 0.72$

and with the 10-fold cross validation, $\tilde{c}_{0.7}^{cv} = 0.72$. Note that $\hat{c}_{0.7} = 0.72$ for $\hat{K}_{0.7}(\cdot)$ presented in Section 2. The distribution of $\tilde{\eta}$ is approximately normal with mean 0 and standard error 0.024. Thus, with 95% chance, the true coverage levels of $\tilde{K}_{0.7}$ and $\tilde{K}_{0.7}^{cv}$ are between 0.65 and 0.74.

4. A NUMERICAL STUDY FOR EXAMINING FINITE SAMPLE PROPERTIES OF $\tilde{K}_\eta(\cdot)$ and $\tilde{K}_\eta^{cv}(\cdot)$

We conducted an extensive simulation study to examine the performance of the empirically calibrated $\tilde{K}_\eta(\cdot)$, the $\mathcal{K}$-fold cross-validated counterpart $\tilde{K}_\eta^{cv}(\cdot)$ and the corresponding conditional set $\hat{I}_\eta(\cdot)$ under various scenarios with small, moderate and large sample sizes.

First, we mimicked the liver surgery study to establish a true model for generating the data $\{(Y_i, Z_i), i = 1, \cdots, n\}$. Specifically, we fitted the observed liver surgery data (n=54) via a location-scale model (2.1) with $h(\cdot)$ being the natural logarithm, $g(\cdot)$ being the identity function and $\sigma(\cdot)$ being the exponential function, but with only two covariates, the standardized prognostic index (PIndex) and the enzyme function test score (EScore). The regression coefficients of this true model are the estimates for β and γ obtained from (2.8) and (2.9), respectively. The true model for our simulation study is

$$\begin{aligned}\log Y = {} & (5.08, 0.38, 0.43)(1, \text{PIndex}, \text{EScore})' + \\ & \exp((-1.32, 0.09, 0.01)(1, \text{PIndex}, \text{EScore})')\ V, \end{aligned} \tag{4.1}$$

where V is normal with mean 0 and variance 1/9.

For a given sample size n, we generated 1000 data sets $\{(Y_i, Z_i), i = 1, \cdots, n\}$ from (4.1). Specifically, for each data set, each of n independent realizations of the covariate vector $Z = (1, \text{BCScore}, \text{PIndex}, \text{EScore}, \text{LScore})'$ was generated from the joint empirical distribution based on the observed 54 covariate vectors, and the corresponding Y was generated from (4.1). The prediction sets, $\hat{I}_{0.8}(\cdot), \tilde{K}_{0.8}(\cdot)$, and $\tilde{K}_{0.8}^{cv}(\cdot)$ were then constructed as described in Sections 2 and 3 under the following six different working models. Note that the error term ϵ in each model has mean 0 and variance one, but its distribution function is unspecified.

(M1) A location-scale model (4.1):

$$\log Y = \beta'(1, \text{PIndex}, \text{EScore})' + \exp\{\gamma'(1, \text{PIndex, EScore})'\}\ \epsilon,$$

(M2) A misspecified location model for $\log Y$:

$$\log Y = \beta'(1, \text{PIndex}, \text{EScore})' + \exp(\gamma)\epsilon.$$

(M3) A misspecified location-scale model for Y:

$$Y = \beta'(1, \text{PIndex}, \text{EScore})' + \exp\{\gamma'(1, \text{PIndex}, \text{EScore})'\}\ \epsilon.$$

(M4) An under-fitted location-scale model for $\log Y$ with a single covariate PIndex:

$$\log Y = \beta'(1, \text{PIndex})' + \exp\{\gamma'(1, \text{PIndex})'\}\epsilon.$$

(M5) An over-fitted location-scale model for $\log Y$ with all four covariates:

$$\log Y = \beta' Z + \exp(\gamma' Z)\ \epsilon.$$

(M6) A misspecified location-scale model for Y^3:

$$Y^3 = \beta'(1, \text{PIndex}, \text{EScore})' + \exp\{\gamma'(1, \text{PIndex}, \text{EScore})'\}\ \epsilon.$$

For each working model, we estimate β and γ via (2.8) and (2.9), and then use (2.10) to obtain the working prediction density function with the smooth parameter selected through cross-validation methods as in the analysis of the liver surgery data in Section 2. For the $\mathcal{K}$-fold cross validation procedure in the simulation study, we let $\mathcal{K} = 5$ for $50 < n < 100$ and let $\mathcal{K} = 10$, for $n \geq 100$. For each realized prediction procedure, the prediction coverage level, PCL, was obtained from (4.1). For each working model and each prediction procedure, there are 1000 estimated realizations of PCL. For each case, we examine closely whether the normal distribution $\text{N}(\eta, \eta(1-\eta)/n)$ is a good approximation to the sampling distribution of PCL based on those 1000 realizations. In Table 2, we report the results with $\eta = 0.8, n = 54, 200, 600$. In the table, each entry under the heading "EPCL$_Y$", the empirical prediction coverage level for the response Y, is the average of the above 1000 realized PCL's. Note that for cases with $n = 54$, $\tilde{K}_{0.8}(\cdot)$ has noticeable downward bias with respect to the prediction coverage level. On the other hand, its cross validated counterpart behaves quite well. For each prediction procedure, the entry under the heading "ESCL$_\eta$, the empirical sampling coverage level for $\tilde{\eta}$ or $\tilde{\eta}^{cv}$, is the proportion of the 1000 realized true coverage levels that belong to the 0.95 two-sided interval $(\eta - 1.96\sqrt{\eta(1-\eta)/n}, \eta + 1.96\sqrt{\eta(1-\eta)/n})$. Although the distribution of the true coverage level for $\hat{I}$ may not be approximated well by a normal, for comparison, we also report the coverage levels for $\hat{I}$ in the table. All ESCL$_\eta$'s are quite close to 0.95 with $\tilde{K}_{0.8}^{cv}(\cdot)$, but not so with $\tilde{K}_{0.8}(\cdot)$. Under the heading "EAS", each entry is the empirical average size based on 1000 realized $\mathrm{E}\|\tilde{K}_{0.8}(Z^0)\|$ or $\mathrm{E}\|\tilde{K}_{0.8}^{cv}(Z^0)\|$ or $\mathrm{E}\|\hat{I}_{0.8}(Z^0)\|$. Although among three prediction procedures, on average $\tilde{K}_{0.8}(\cdot)$ is the smallest for all cases considered here, unfortunately it may not have the desirable prediction coverage level. The empirically calibrated, cross validated procedure $\tilde{K}_{0.8}^{cv}(\cdot)$ has correct coverage level and also produces uniformly smaller regions than $\hat{I}_{0.8}(\cdot)$ across all models studied here.

We also examined the situation that the working prediction density function is based on a parametric model. For example, we let the error term ϵ in each of the working models be the standard normal or equivalently let $\hat{f}_0$ in (2.10) be the density function of the standard normal. In Table 3, we report the results obtained under the same setting as Table 2, but with this specific parametric modeling. The $\hat{I}_{0.80}(\cdot)$ does not perform well at all with respect to the prediction coverage level and average length. On the other hand, the cross validated $\tilde{K}_{0.8}^{cv}(\cdot)$ continues performing exceptionally well for every case.

5. REMARKS

Based on the results from the extensive numerical study in Section 4, we find that the empirically calibrated prediction procedure $\tilde{K}_\eta(\cdot)$ tends to be too liberal, that is, its prediction coverage probability can be markedly smaller than the nominal level even with moderate sample sizes. On the other hand, the $\mathcal{K}$-fold cross validation procedure performs quite well. Moreover, the extra computation burden for constructing its prediction regions is minimal. We recommend its usage in practice.

If the fitted model is correctly specified, one may construct optimal prediction intervals based on either $\hat{K}_\eta$ or $\tilde{K}_\eta^{cv}$. However, unlike the case for $\tilde{K}_\eta^{cv}$, the limiting distribution of the true coverage level for $\hat{K}_\eta$ may depend on the underlying error distribution.

It is important to note that when the fitted model is not correctly specified, $\tilde{K}_\eta^{cv}(\cdot)$ may not be the optimal prediction set among all the empirically calibrated, cross validated, unconditional prediction procedures. It would be interesting to explore whether one can identify a prediction procedure which produces the smallest size on average over a population of covariate vectors of interest without assuming that the fitted model is correctly specified.

In survival analysis, the response variable is the time to a certain event, which is possibly right-censored. Therefore, the right tail of the prediction density function $f(\cdot; z)$ may not be estimated well semi-parametrically. Moreover, it is not clear how to do the empirical calibration due to the fact that Y may be incompletely observed. It is interesting to explore how to predict future responses when the survival model may be misspecified.

APPENDIX A. PROOF OF OPTIMALITY FOR $\hat{K}_\eta(\cdot)$

First, consider the case that Y is continuous. Assume that $\hat{\varepsilon}_1 = \sup_{y,z} |\hat{f}(y; z) - f(y; z)| \to 0$ in probability, as $n \to \infty$. Then,

$$|\mathrm{pr}\{Y^0 \in \hat{L}_\eta(Z^0)\} - \eta| \leq \int_{\mathcal{D}} \int |f(y; z) - \hat{f}(y; z)| dy dH(z) \to 0,$$

in probability and the true average coverage level of $\hat{L}_\eta(\cdot)$ converges to η in probability, as $n \to \infty$. This is trivially true when Y assumes a finite number of possible values.

To demonstrate $\hat{K}_\eta(\cdot)$ is optimal, again, assume that Y is continuous first. Let $K_\eta(z) = \{y : f(y; z) \geq c_\eta\}$, where c_η is the solution to $\eta_0(c) \equiv \mathrm{pr}(f(Y^0; Z^0) \geq c) = \eta$. Assuming that $\hat{\varepsilon}_1 \to 0$ in probability and $\eta_0(\cdot)$ has a nonzero derivative at c_η , in the first step we show that

$$\sup_z \left\| \hat{K}_\eta(z) - K_\eta(z) \right\| \equiv \sup_z \int_{-\infty}^{\infty} \left| \mathcal{I}\{\hat{f}(y; z) \geq \hat{c}_\eta\} - \mathcal{I}\left\{f(y; z) \geq c_\eta\right\} \right| dy \to 0$$

in probability, where $\mathcal{I}(\cdot)$ is the indicator function. To this end, we let $\hat{\eta}(c) = \int \mathcal{I}\{\hat{f}(y;z) \geq c\}\hat{f}(y,z)dydH(z)$. Then

$$|\hat{\eta}(c) - \eta_0(c)| \leq \text{pr}\left\{|f(Y^0;Z^0) - c| \leq \hat{\varepsilon}_1\right\} + \int_{\mathcal{D}}\int |\hat{f}(y;z) - f(y;z)|dydH(z) \to 0,$$

in probability. Since $\hat{\eta}(c)$ is monotone decreasing, it follows that $\sup_c |\hat{\eta}(c) - \eta_0(c)| \to 0$ in probability and thus $\hat{c}_\eta$ is consistent for c_η. This, together with the uniform consistency of $\hat{f}(y;z)$, implies that

$$\begin{aligned}
&\sup_z \|\hat{K}_\eta(z) - K_\eta(z)\| \\
&\leq \sup_z \int_{-\infty}^{\infty} \Big[\mathcal{I}\{c_\eta > f(y;z) \geq \hat{c}_\eta - \hat{\varepsilon}_1\} + \mathcal{I}\{\hat{c}_\eta + \hat{\varepsilon}_1 \geq f(y;z) \geq c_\eta\}\Big]dy \\
&\leq \sup_z \int_{-\infty}^{\infty} \mathcal{I}(|f(y;z) - c_\eta| \leq \hat{\varepsilon}_1 + |\hat{c}_\eta - c_\eta|)dy. \qquad \text{(A.1)}
\end{aligned}$$

Furthermore, if there exists a positive integer m such that $f(y,z)$ has continuous partial derivatives in y up to order m and $\inf_{\{(y,z):f(y;z)=c_\eta\}} \max\{|\partial f(y,z)/\partial y|, \cdots, |\partial^m f(y,z)/\partial y^m|\} > 0$, then (A.1) $\to 0$ in probability. By the triangle inequality, it implies that $\text{E}\|\hat{K}_\eta(Z^0)\|$ converges to $\text{E}\|K_\eta(Z^0)\|$ in probability, where the expectation is with respect to Z^0.

Lastly, we show that the expected length of $K_\eta(\cdot)$ is the shortest among all $L_\eta(\cdot)$. Without loss of any generality, we only consider $L_\eta(Z^0) = \{y : f(y;Z^0) \geq c(Z^0)\}$ with $c(\cdot)$ possibly covariate dependent and $\text{pr}\{f(Y^0;Z^0) \geq c(Z^0)\} = \eta$. It is equivalent to show that for any given $L_\eta(Z^0)$, if there exists a constant c^* such that the prediction sets $K_\eta(Z^0) = \{y : f(y;Z^0) \geq c^*\}$ have the same expected length as that of $L_\eta(Z^0)$, then $\text{pr}\{f(Y^0;Z^0) \geq c(Z^0)\} \leq \text{pr}\{f(Y^0;Z^0) \geq c^*\}$. This follows from the fact that

$$\begin{aligned}
&\text{pr}\{f(Y^0,Z^0) \geq c(Z^0)\} - \text{pr}\{f(Y^0;Z^0) \geq c^*\} \\
&= \text{E}\left[\int_{c^*>f(y;Z^0)\geq c(Z^0)} f(y;Z^0)dy - \int_{c(Z^0)>f(y;Z^0)\geq c^*} f(y;Z^0)dy\right] \\
&\leq c^*\text{E}\left[\int_{c^*>f(y;Z^0)\geq c(Z^0)} dy - \int_{c(Z^0)>f(y;Z^0)\geq c^*} dy\right] \\
&= c^*\text{E}\left[\int_{f(y;Z^0)\geq c(Z^0)} dy - \int_{f(y;Z^0)\geq c^*} dy\right] = 0.
\end{aligned}$$

The proof for the case that Y has a finite number of possible values is straightforward.

APPENDIX B. PROOF OF CONSISTENCY FOR $\hat{\Theta}$

Here we show that $\hat{\Theta} = \{\hat{\theta}, \hat{f}_0(\cdot)\}$ is consistent for $\Theta_0 = \{\theta_0 = (\beta_0', \gamma_0')', f_0(\cdot)\}$, where β_0 is the solution to the equation $E[Z\{h(Y) - g(\beta' Z)\}] = 0$, γ_0 is the solution to the equation $E(Z[\{h(Y) - g(\beta_0' Z)\}^2 - \sigma(\gamma' Z)^2])$, and $f_0(\cdot)$ is the true density function of $\{h(Y) - g(\beta_0' Z)\}/\sigma(\gamma_0' Z)$. Note that when Model (2.1) holds, β_0 and γ_0 are the true regression parameters and f_0 is the true density function of ϵ.

First, it follows from the same argument as given in Tian et al (2006) that θ_0 exists and is unique. The consistency of $\hat{\beta}$ for β_0 follows directly from Tian et al (2006). This, together with the consistency of $\hat{\beta}$ and the standard M-estimation theory (van der Vaart, 1998, Chapter 5) that $\hat{\gamma}$ is consistent for γ_0. Furthermore, it is straightforward to show that $\hat{\theta} - \theta_0 = O_p(n^{-1/2})$, where $\hat{\theta} = (\hat{\beta}', \hat{\gamma}')'$. The consistency of $\hat{f}(y; z)$ for the binary case follows directly from the consistency of $\hat{\theta}$. Now, for the location-scale working model (2.1), let

$$f(y; z) = \frac{\dot{h}(y)}{\sigma(\gamma_0' z)} f_0 \left\{ \frac{h(y) - g(\beta_0' z)}{\sigma(\gamma_0' z)} \right\}.$$

Here we assume that $f_0(\cdot)$, $\sigma(\cdot)$ and $g(\cdot)$ are continuously differentiable and $\inf_{z \in \mathcal{D}} \{\sigma(\gamma_0' z)\} > 0$. Since $\hat{\theta}$ is consistent for θ_0, the uniform consistency of $\hat{f}(y; z)$ for $f(y; z)$ holds if $\hat{f}_0(x)$ is uniformly consistent for $f_0(x)$. Note that $\sup_x |\hat{f}_0(x) - f_0(x)|$ is bounded by

$$\begin{aligned}
&\sup_x \left| \int \phi_\tau(u - x) d\{\hat{H}(u; \hat{\theta}) - H_0(u)\} \right| + \sup_x \left| \int_{-\infty}^{\infty} \phi_\tau(u - x) dH_0(u) - f_0(x) \right| \\
\le &\sup_x \left| \tau^{-2} \int_{-\infty}^{\infty} \dot{\phi}\{(u - x)/\tau\} \{\hat{H}(u; \hat{\theta}) - H_0(u)\} du \right| + O(\tau^2) \\
\le &\sup_u \left| \hat{H}(u; \hat{\theta}) - H_0(u) \right| \tau^{-1} \int |\dot{\phi}(u)| du + O(\tau^2)
\end{aligned}$$

where $\hat{H}(u; \hat{\theta}) = n^{-1} \sum_{i=1}^n \mathcal{I}(\hat{e}_i \le u)$ and $H_0(u) = \mathrm{pr}(e_i \le u)$. It is not difficult to show that the set of functions indexed by (β, γ, u): $\{\mathcal{I}[\{h(y) - g(\beta' z)\}/\sigma(\gamma' z) \le u] : |\beta - \beta_0| + |\gamma - \gamma_0| \le \delta, u \in R\}$ is VC class for a small positive δ. This, coupled with the fact that $n^{1/2}(\hat{\theta} - \theta_0)$ converges weakly to a mean zero multivariate normal, implies that $n^{1/2}\{\hat{H}(u;\hat{\theta})-H_0(u)\}$ converges weakly to a tight, mean zero Gaussian process in u. Therefore, $\sup_x|\hat{f}_0(x)-f_0(x)| = O_p(\tau^{-1}n^{-1/2}+\tau^2) = o_p(1)$, when $\tau = o_p(1)$ and $n^{1/2}\tau \to \infty$.

APPENDIX C. PROOF OF LARGE SAMPLE PROPERTIES FOR $\tilde{K}_\eta(\cdot)$

First we establish the uniform consistency of $\hat{\eta}(c) = n^{-1}\sum_{i=1}^n \mathcal{I}\{\hat{f}(Y_i;Z_i)\geq c\}$ to $\eta_0(c)$. The consistency of $\tilde{c}_\eta$ to c_η follows from the same arguments in Appendix A. To this end,

$$\begin{aligned}|\hat{\eta}(c)-\eta_0(c)| &\leq n^{-1}\sum_{i=1}^n \mathcal{I}\left\{|f(Y_i;Z_i)-c|\leq\hat{\varepsilon}_1\right\}+\hat{\varepsilon}_2\\ &\leq 3\hat{\varepsilon}_2+\eta_0(c-\hat{\varepsilon}_1)-\eta_0(c+\hat{\varepsilon}_1)\to 0, \qquad \text{(C.1)}\end{aligned}$$

in probability, where $\hat{\varepsilon}_2 = \sup_c|n^{-1}\sum_{i=1}^n \mathcal{I}\{f(Y_i;Z_i)\geq c\}-\eta_0(c)| = o_p(1)$. The consistency of $\tilde{\eta}$ follows directly from the continuous mapping theorem, the uniform consistency of $\hat{f}(y;z)$ and the consistency of $\tilde{c}_\eta$.

To derive the large sample distribution for $n^{1/2}(\tilde{\eta}-\eta)$, we need the convergence rate for $\hat{f}(y;z)$ and $\tilde{c}_\eta$. By a functional central limit theorem (Pollard, 1990), $n^{-1/2}[\sum_{i=1}^n \mathcal{I}\{f(Y_i;Z_i)\geq c\}-\eta_0(c)]$ converges weakly to a mean zero Gaussian process. This, coupled with $\hat{\theta}-\theta_0 = O_p(n^{-1/2})$ and $\sup_x|\hat{f}_0(x)-f_0(x)| = O_p(\tau^{-1}n^{-1/2}+\tau^2)$, implies that

$$\hat{\varepsilon}\equiv\sup_{y;z}\left|\hat{f}(y;z)-f(y;z)\right|+\sup_c\left|n^{-1}\sum_{i=1}^n\mathcal{I}\left\{f(Y_i;Z_i)\geq c\right\}-\eta_0(c)\right| = O_p(\tau^{-1}n^{-1/2}+\tau^2). \quad (4{\cdot}1)$$

It then follows from (C.1) that $\tilde{c}_\eta - c_\eta = O_p(\tau^{-1}n^{-1/2}+\tau^2)$. Now, let

$$\hat{M}(c,f) = n^{-1/2}\sum_{i=1}^n\left\{\mathcal{I}(f(Y_i;Z_i)\geq c)-\eta_0(c,f)\right\}, \quad \text{where} \quad \eta_0(c,f) = \text{pr}(f(Y,Z)\geq c).$$

If we can show that

$$\hat{R}_0 = \hat{M}(\tilde{c}_\eta,\hat{f})-\hat{M}(c_\eta,f)\to 0, \qquad \text{in probability,} \qquad \text{(C.2)}$$

then

$$n^{1/2}(\tilde{\eta}-\eta) = n^{-1/2}\sum_{i=1}^{n}\left\{\eta_0(\tilde{c}_\eta,\hat{f}) - \mathcal{I}(\hat{f}(Y_i;Z_i)\geq \tilde{c}_\eta)\right\}$$
$$= -n^{-1/2}\sum_{i=1}^{n}\{\mathcal{I}(f(Y_i;Z_i)\geq c_\eta)-\eta\} + o_p(1).$$

Therefore, by the central limit theorem, we have $n^{1/2}(\tilde{\eta}-\eta) \overset{\mathcal{D}}{\rightarrow} N(0,\eta(1-\eta))$. We now prove (C.2) for the location-scale working model. The binary case can shown easily using similar arguments. To this end, we let $\theta = (\beta',\gamma')'$

$$e_\theta(Y_i;Z_i) = \frac{h(Y_i)-g(\beta'Z_i)}{\sigma(\gamma'Z_i)}, \qquad \hat{G}_\theta(e,s) = n^{-1}\sum_{i=1}^{n}\mathcal{I}\left\{e_\theta(Y_i;Z_i)\leq e, \frac{\sigma(\gamma'Z_i)}{\dot{h}(Y_i)}\leq s\right\},$$
$$G_\theta(e,s) = \text{pr}\left\{e_\theta(Y_i;Z_i)\leq e, \frac{\sigma(\gamma'Z_i)}{\dot{h}(Y_i)}\leq s\right\}, \quad \text{and} \quad g(e,s) = \frac{\partial^2 G_{\theta_0}(e,s)}{\partial e\partial s},$$

where the density function $g(e,s)$ is assumed to have bounded continuous derivatives up to the second order. Then

$$\eta_0(c,f) = \text{pr}\left\{f_0(e_\theta(Y,Z))\geq c\frac{\sigma_\gamma(Z)}{\dot{h}(Y)}\right\} = \int \mathcal{I}(f_0(e)\geq cs)dG_\theta(e,s)$$

and

$$\hat{R}_0 = n^{1/2}\int\int_{-\infty}^{\infty}\mathcal{I}\left\{\hat{f}_0(e)\geq\tilde{c}_\eta s\right\}d\left\{\hat{G}_{\hat{\theta}}(e,s)-G_{\hat{\theta}}(e,s)\right\}$$
$$-\, n^{1/2}\int\int_{-\infty}^{\infty}\mathcal{I}\left\{f_0(e)\geq c_\eta s\right\}d\left\{\hat{G}_{\theta_0}(e,s)-G_{\theta_0}(e,s)\right\}.$$

By the standard empirical processes theory (Pollard, 1990), $n^{1/2}\{\hat{G}_\theta(e,s)-G_\theta(e,s)\}$ converges weakly to a mean-zero Gaussian process in (θ,e,s), and thus is equi-continuous. Now, let

$$\hat{R} = n^{1/2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\mathcal{I}\left\{\hat{f}_0(e)\geq\tilde{c}_\eta s\right\}-\mathcal{I}\left\{f_0(e)\geq c_\eta s\right\}\right]d\left\{\hat{G}_{\theta_0}(e,s)-G_{\theta_0}(e,s)\right\}.$$

Given the consistency of $\hat{\theta}$, it follows that $\hat{R}_0-\hat{R}=o_p(1)$ and thus to show (C.2), it suffices to show that $\hat{R}=o_p(1)$. To this end, let $\hat{R}=\hat{R}_1+\hat{R}_2$, where

$$\hat{R}_1 = \int\int_{-\infty}^{\infty}\left[\mathcal{I}\left\{\hat{f}_0(e)\geq\tilde{c}_\eta s\right\}-\mathcal{I}\left\{f_0(e)\geq c_\eta s\right\}\right]d\left[n^{1/2}\left\{\hat{G}_{\theta_0}(e,s)-\hat{G}(e,s)\right\}\right],$$
$$\hat{R}_2 = n^{1/2}\int\int_{-\infty}^{\infty}\left[\mathcal{I}\left\{\hat{f}_0(e)\geq\tilde{c}_\eta s\right\}-\mathcal{I}\left\{f_0(e)\geq c_\eta s\right\}\right]\{\hat{g}(e,s)-g(e,s)\}dsde,$$

$\hat{g}(e,s) = n^{-1}\sum_{i=1}^{n}\phi_{n^{-v}}\{e_{\theta_0}(Y_i;Z_i) - e\}\phi_{n^{-v}}\{\sigma(\gamma_0'Z_i)/\dot{h}(Y_i) - s\}, 1/4 < v < 1/2$, and $\hat{G}(e,s) = \int_{-\infty}^{e}\int_{-\infty}^{s}\hat{g}(a,b)dadb$. It follows from the asymptotic properties of the bivariate kernel density estimator (Rosenblatt, 1976; Silverman, 1986) and that $|\tilde{c}_\eta - c_\eta| + \sup_x|\hat{f}_0(x) - f_0(x)| = O_p(\tau^{-1}n^{-1/2} + \tau^2)$, we have

$$\sup_{e,s}|\hat{g}(e,s) - g(e,s)| = o_p(n^{\epsilon+v-1/2}), \text{forall}\epsilon > 0, \text{and}$$
$$\hat{\varepsilon}_3 = \tilde{c}_\eta^{-1}\int_{-\infty}^{\infty}\left|\hat{f}_0(e) - f_0(e)\right|de + |\tilde{c}_\eta^{-1} - c_\eta^{-1}| = O_p(\tau^{-1}n^{-1/2} + \tau^2).$$

Lastly, it follows from van der Vaart (1994), the "smoothed" empirical process $\hat{G}(e,s)$ is asymptotically equivalent to $\hat{G}_{\theta_0}(e,s)$, i.e., $\sup_{e,s}\left|\hat{G}_{\theta_0}(e,s) - \hat{G}(e,s)\right| = o_p(n^{-1/2})$. These, together with Lemma 1 of Bilias et al (1997), imply that $\hat{R}_1 = o_p(1)$ and

$$\begin{aligned}\hat{R}_2 &\le n^{1/2}\sup_{e,s}|\hat{g}(e,s) - g(e,s)|\int\int_{-\infty}^{\infty}\left|\mathcal{I}\left\{\hat{f}_0(e) \ge \tilde{c}_\eta s\right\} - \mathcal{I}\left\{f_0(e) \ge c_\eta s\right\}\right|dsde\\ &\le n^{1/2}\sup_{e,s}|\hat{g}(e,s) - g(e,s)|\,\hat{\varepsilon}_3 = O_p(\tau^{-1}n^{\epsilon+v-1/2} + \tau^2 n^{\epsilon+v}).\end{aligned}$$

Therefore, if $\tau = O_p(n^{-\delta})$ and $v \to 1/4$, for $1/8 < \delta < 1/4$, $\hat{R} \to 0$ in probability and consequently $n^{1/2}(\tilde{\eta} - \eta)$ converges in distribution to a normal with mean 0 and variance $\eta(1-\eta)$.

APPENDIX D. PROOF OF LARGE SAMPLE PROPERTIES FOR $\tilde{K}_\eta^{cv}(\cdot)$

We first show that $\tilde{c}_\eta^{cv}$ is consistent for c_η for any given η. Note that $\tilde{c}_\eta^{cv}$ is the solution to $\tilde{\eta}_{cv}(c) = \eta$, where

$$\tilde{\eta}_{cv}(c) \equiv n^{-1}\sum_{k=1}^{\mathcal{K}}\sum_{\{i:\xi_i=k\}}\mathcal{I}\left\{\hat{f}_{(-k)}(Y_i;Z_i) \ge c\right\}.$$

Using the arguments in Appendix C, we can show that

$$\hat{\varepsilon}_{cv} \equiv \sum_{k=1}^{K}\sup_{y;z}\left|\hat{f}_{(-k)}(y;z) - f(y;z)\right| + \sup_c\left|n^{-1}\sum_{i=1}^{n}\mathcal{I}\left\{f(Y_i;Z_i) \ge c\right\} - \eta_0(c)\right| = O_p(n^{-1/2}\tau^{-1} + \tau^2).$$

It follows that $|\tilde{\eta}_{cv}(c) - \eta_0(c)|$ is bounded by

$$\begin{aligned}&n^{-1}\sum_{k=1}^{K}\sum_{\xi_i=k}\left[\mathcal{I}\left\{\hat{f}_{(-k)}(Y_i;Z_i) \ge c, f(Y_i;Z_i) < c\right\} + \mathcal{I}\left\{\hat{f}_{(-k)}(Y_i;Z_i) < c, f(Y_i;Z_i) \ge c\right\}\right] + \hat{\epsilon}_2\\ &\le n^{-1}\sum_{i=1}^{n}\mathcal{I}\left\{c + \hat{\varepsilon}_{cv} \ge f(Y_i;Z_i) \ge c - \hat{\varepsilon}_{cv}\right\} + \hat{\varepsilon}_2 \le 3\hat{\varepsilon}_2 + \eta_0(c - \hat{\varepsilon}_{cv}) - \eta_0(c + \hat{\varepsilon}_{cv}).\end{aligned}$$

Thus, $\sup_c |\tilde{\eta}_{cv}(c) - \eta_0(c)| = O_p(n^{-1/2}\tau^{-1} + \tau^2)$. Given the assumption that $\eta_0(c)$ has a nonzero derivative at $c = c_\eta$, we have $|\tilde{c}_\eta^{cv} - c_\eta| = O_p(\tau^{-1}n^{-1/2} + \tau^2)$ and therefore, $\tilde{c}_\eta^{cv}$ is consistent for c_η. It then follows from the same argument as given in Appendix C that $\sup_z \|\tilde{K}_\eta^{cv}(z) - K_\eta(z)\| \to 0$, in probability. The weak convergence of $n^{1/2}(\tilde{\eta}^{cv} - \eta)$ to $N(0, \eta(1-\eta))$ follows directly from the convergence rate of $\tilde{c}_\eta^{cv}$ and the same arguments as given in Appendix C.

References

Bayarri, M. J. & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science* **19**, 58–80.

Bilias, Y., Gu, M. & Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. *The Annals of Statistics* **25**, 662–682.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–360.

Box, G. & Tiao, G. (1973). *Bayesian Inference in Statistical Analysis.* Addision-Wesley, Reading Mass.

Carroll, R. J. & Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman & Hall Ltd.

Carroll, R. J. & Ruppert, D. (1991). Prediction and tolerance intervals with transformation and/or weighting. *Technometrics* **33**, 197–210.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**, 461–470.

Gail, M. H., Wieand, S. & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71**, 431–444.

Geisser, S. (1993). *Predictive Inference: an Introduction.* Chapman & Hall Ltd.

Lagakos, S. W. & Schoenfeld, D. A. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* **40**, 1037–1048.

Lin, D. Y. & Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84**, 1074–1078.

Lin, D. Y. & Wei, L. J. (1991). Goodness-of-fit tests for the general Cox regression model. *Statistica Sinica* **1**, 1–17.

Manes, C., Pfeffer, M., Rutherford, J., Greaves, S., Rouleau, J., Arnold, J., Menapace, F. & Solomon, S. (2003). Value of the electrocardiogram in predicting left ventricular enlargement and dysfunction after myocardial infarction. *Am. J. Med* **114**, 99–105.

Neter, J., Wasserman, W. & Kutner, M. H. (1985). *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs.* Richard D. Irwin Inc.

Neyman, J. (1977). Frequentist probability and frequent statistics. *Synthese: An International Journal for Epistemology, Methodology and Philosophy of Science* **36**, 97–132.

Olive, D. (2006). Prediction intervals for regression models. *Computational Statistics and Data Analysis,* to appear.

Patel, J. K. (1989). Prediction intervals: A review. *Communications in Statistics: Theory and Methods* **18**, 2393–2465.

Pfeffer, M., Greaves, S., Arnold, J., Glynn, R., LaMotte, F., Lee, R., Menapace, F., Rapaport, E., Ridker, P., Rouleau, J., Solomon, S. & Hennekens, C. (1997). Early versus delayed angiotensin-converting enzyme inhibition therapy in acute myocardial infarction. the healing and early afterload reducing therapy trial. *Circulation* **97**, 2643–51.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* Institute of Mathematical Statistics.

Preston, S. (2000). Teaching prediction intervals. *Journal of Statistics Education* **8**, 1–1.

Rosenblatt, M. (1976). On the maximal deviation of k-dimensional density estimates. *The Annals of Probability* **4**, 1009–1015.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12**, 1151–1172.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9**, 65–78.

Schmoyer, R. L. (1992). Asymptotically valid prediction intervals for linear models. *Technometrics* **34**, 399–408.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall Ltd.

Stine, R. A. (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association* **80**, 1026–1031.

Struthers, C. A. & Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73**, 363–369.

Tian, L., Cai, T., Goetghebeur, E. & Wei, L. J. (2006). Model evaluation based on the distribution of estimated absolute prediction error. *Biometrika,* in revision.

van der Vaart, A. W. (1994). Weak convergence of smoothed empirical processes. *Scandinavian Journal of Statistics* **21**, 501–504.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Table 1. *Estimates of the regression parameter (standard errors) of models for liver surgery survival data and HEART Trial data*

(a) Liver Surgery Study

	Intercept	BCScore	PIndex	EScore	LScore
$\hat{\beta}$	1.125(0.113)	0.158(0.012)	0.021(0.001)	0.022(0.001)	0.004(0.018)
$\hat{\gamma}$	-1.272(1.039)	-0.190(0.114)	0.003(0.006)	0.000(0.006)	-0.088(0.141)

(b) HEART Trial

	Intercept	EFrac	AveST	MaxST
$\hat{\beta}$	2.564(1.079)	-0.087(0.019)	-1.516(0.725)	0.993(0.392)

Table 2. *Summary of finite sample properties for* $\tilde{K}_{0.8}(\cdot)$, $\tilde{K}^{cv}_{0.8}(\cdot)$ *and* $\hat{I}_{0.8}(\cdot)$ *for various working models with unspecified error distribution functions*

		EPCL$_Y$[1]			ESCL$_\eta$[2]			EAS[3]		
n	Model	$\tilde{K}_{0.8}$	$\tilde{K}^{cv}_{0.8}$	$\hat{I}_{0.8}$	$\tilde{K}_{0.8}$	$\tilde{K}^{cv}_{0.8}$	$\hat{I}_{0.8}$	$\tilde{K}_{0.8}$	$\tilde{K}^{cv}_{0.8}$	$\hat{I}_{0.8}$
54	M1	.74	.80	.79	.78	.93	.94	35	41	47
	M2	.75	.79	.80	.82	.94	.95	36	41	46
	M3	.74	.80	.80	.79	.93	.96	69	84	95
	M4	.74	.80	.78	.76	.92	.92	147	174	189
	M5	.72	.80	.77	.68	.92	.87	34	43	46
	M5	.74	.80	.63	.80	.95	.23	348	399	239
200	M1	.78	.80	.81	.89	.95	.94	38	39	46
	M2	.79	.80	.81	.90	.94	.94	38	39	44
	M3	.78	.80	.81	.86	.92	.97	76	77	89
	M4	.78	.80	.79	.84	.93	.94	166	169	194
	M5	.78	.80	.81	.86	.94	.96	39	39	46
	M6	.79	.80	.70	.91	.92	.19	445	463	312
600	M1	.79	.80	.81	.91	.94	.90	38	39	46
	M2	.79	.80	.81	.91	.94	.89	38	39	45
	M3	.79	.80	.80	.91	.94	.96	76	77	88
	M4	.79	.80	.80	.88	.94	.94	164	169	192
	M5	.79	.80	.81	.90	.94	.92	38	39	46
	M5	.79	.80	.71	.94	.94	.01	482	491	313

[1]: the empirical prediction coverage level for Y

[2]: the empirical sampling coverage level of the true prediction coverage probability

[3]: the empirical average size

Table 3. *Summary of finite sample properties for* $\tilde{K}_{0.8}(\cdot)$, $\tilde{K}^{cv}_{0.8}(\cdot)$ *and* $\hat{I}_{0.8}(\cdot)$ *for various working models with normal error distribution functions*

		EPCL$_Y$[1]			ESCL$_\eta$[2]			EAS[3]		
n	Model	$\tilde{K}_{0.8}$	$\tilde{K}^{cv}_{0.8}$	$\hat{I}_{0.8}$	$\tilde{K}_{0.8}$	$\tilde{K}^{cv}_{0.8}$	$\hat{I}_{0.8}$	$\tilde{K}_{0.8}$	$\tilde{K}^{cv}_{0.8}$	$\hat{I}_{0.8}$
54	M1	.76	.80	.76	.88	.95	.89	37	41	43
	M2	.77	.79	.77	.90	.94	.93	37	40	42
	M3	.77	.80	.83	.91	.95	.95	85	93	108
	M4	.76	.80	.74	.89	.94	.81	159	173	168
	M5	.74	.80	.73	.78	.93	.76	35	43	42
	M6	.77	.80	.51	.90	.95	.13	330	365	236
200	M1	.79	.80	.79	.93	.95	.97	38	39	44
	M2	.79	.80	.79	.93	.95	.98	38	39	43
	M3	.79	.80	.86	.92	.94	.37	85	86	114
	M4	.79	.80	.75	.93	.93	.67	168	171	175
	M5	.78	.80	.78	.91	.96	.93	38	39	44
	M6	.79	.80	.47	.93	.95	.01	322	333	227
600	M1	.80	.80	.80	.94	.94	.98	38	39	45
	M2	.80	.80	.80	.94	.94	.98	39	39	43
	M3	.80	.80	.87	.94	.94	.01	84	85	115
	M4	.80	.80	.78	.93	.94	.30	170	172	177
	M5	.79	.80	.79	.94	.95	.97	38	39	45
	M6	.80	.80	.46	.94	.94	.00	295	298	226

[1]: the empirical prediction coverage level for Y

[2]: the empirical sampling coverage level of the true prediction coverage probability

[3]: the empirical average size

Figure 1. Prediction density function estimates for the survival time with (a) $z^0 = (14.8, 86, 101, 4.1)'$ and (b) $z^0 = (1, 6.6, 77, 46, 1.95)'$.

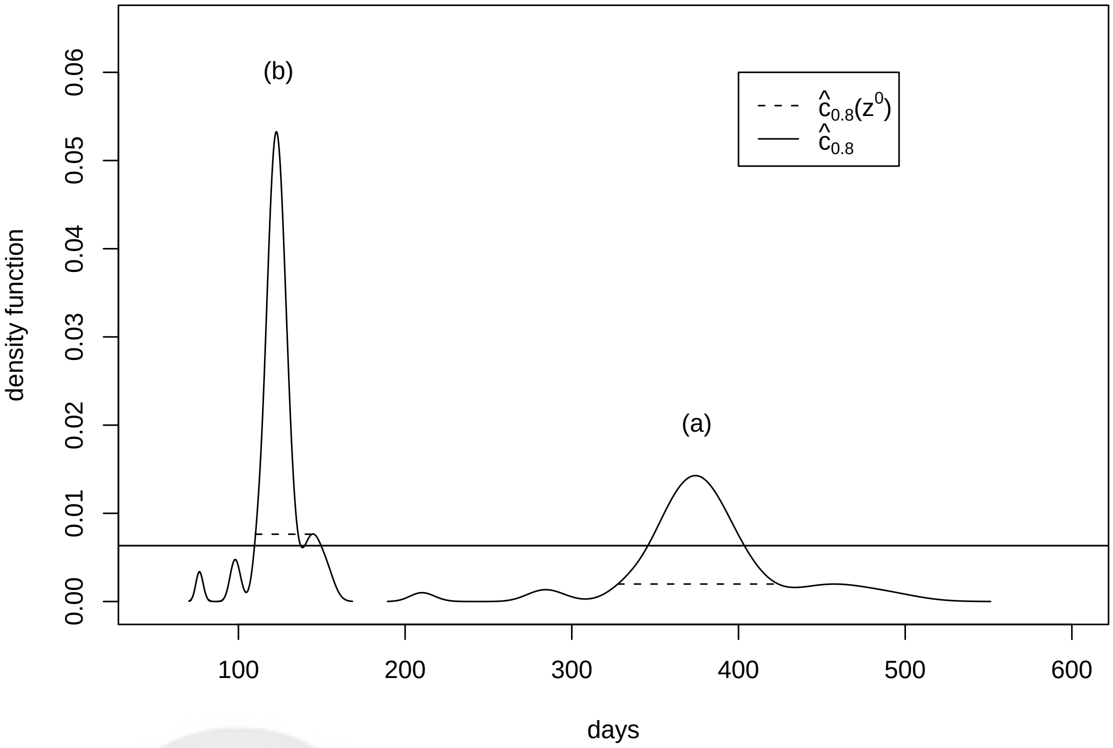
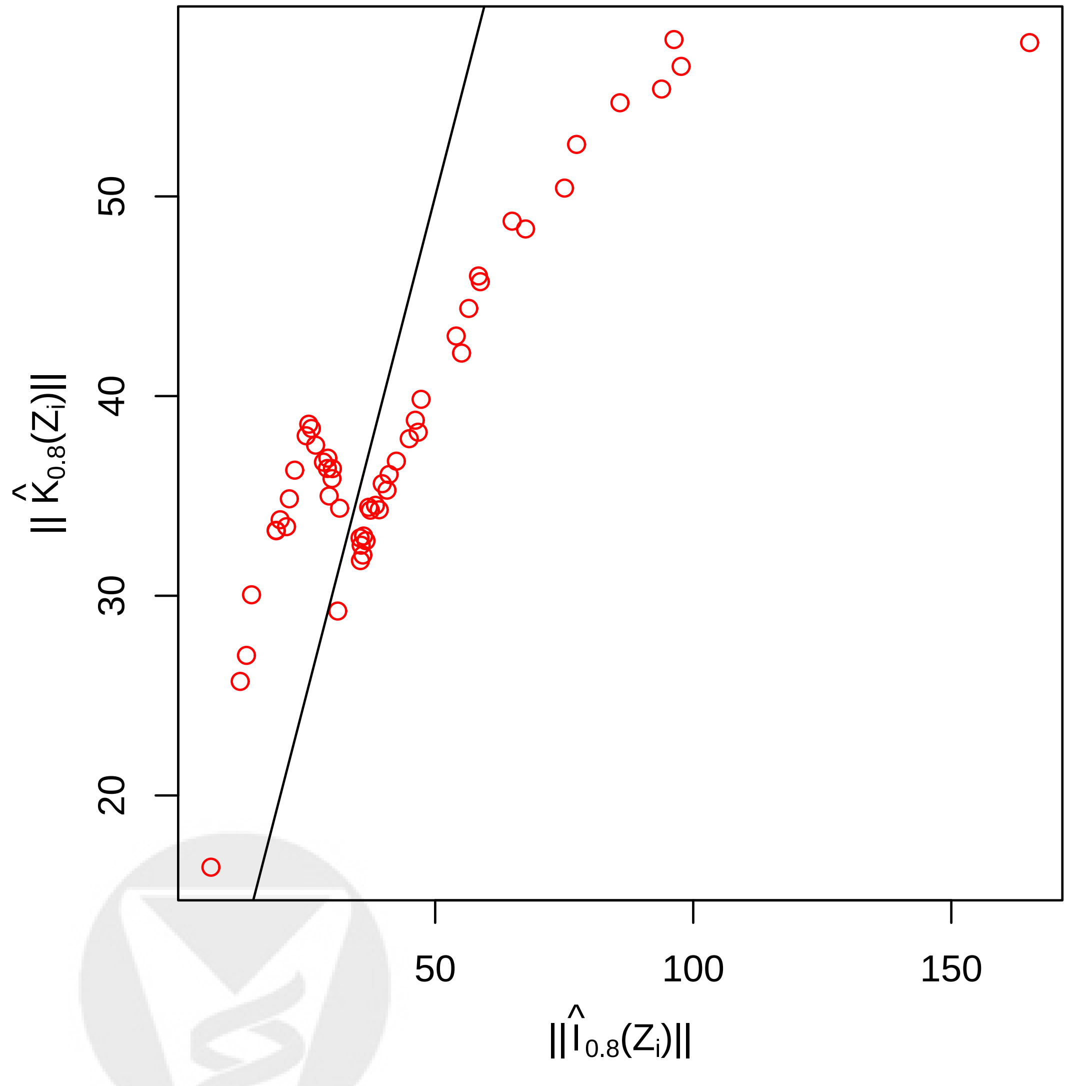


Figure 2. Comparisons between $\hat{K}_{0.8}(\cdot)$ and $\hat{I}_{0.8}(\cdot)$ with respect to the length of the prediction region for the liver surgery survival time data (the solid line is the 45° reference line).



20
30
40
50
$\|\hat{K}_{0.8}(Z_i)\|$
50
100
150
$\|\hat{I}_{0.8}(Z_i)\|$