Prediction Performance of Survival Models

by

Yan Yuan

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2008

Author's Declaration Page

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Statistical models are often used for the prediction of future random variables. There are two types of prediction, point prediction and probabilistic prediction. The prediction accuracy is quantified by performance measures, which are typically based on loss functions. We study the estimators of these performance measures, the prediction error and performance scores, for point and probabilistic predictors, respectively. The focus of this thesis is to assess the prediction performance of survival models that analyze censored survival times. To accommodate censoring, we extend the inverse probability censoring weighting (IPCW) method, thus arbitrary loss functions can be handled. We also develop confidence interval procedures for these performance measures.

We compare model-based, apparent loss based and cross-validation estimators of prediction error under model misspecification and variable selection, for absolute relative error loss (in chapter 3) and misclassification error loss (in chapter 4). Simulation results indicate that cross-validation procedures typically produce reliable point estimates and confidence intervals, whereas model-based estimates are often sensitive to model misspecification. The methods are illustrated for two medical contexts in chapter 5. The apparent loss based and cross-validation estimators of performance scores for probabilistic predictor are discussed and illustrated with an example in chapter 6. We also make connections for performance.

Acknowledgements

It is a long journey and I would like to thank a number of people, without whom my PhD career would not be so enjoyable and memorable.

First and most importantly, I want to thank my thesis advisor, Dr. Jerry Lawless, for his strategic insight, guidance, patience and confidence in me. I feel extremely grateful to get my PhD training with Jerry, from whom I learned tremendously, statistics and beyond, through our communications over the years. He is the best possible advisor who not only educates, but also cares about students.

I wish to thank Dr. Mu Zhu and Dr. Richard Cook who are the department members of my thesis committee. They always take time to answer my questions despite their busy schedule and have given valuable inputs on my research.

I would like to thank Dr. Patricia O'Brien (School of Accounting, University of Waterloo) and Dr. Michal Abrahamowicz (McGill Univeristy Health Centre, Royal Victoria Hospital), the other two members of my examine committee, for their dedication in reading my thesis and for their helpful comments and suggestions.

Many thanks go to Ker-Ai Lee, who advises me whenever I have questions in programming. She is also a good friend, always there when I need support.

Next, I wish to thank Drs. Sunny Wang, Wanhua Su, JiEun Choi, Leilei Zeng, Cindy Fu, Hui Shen, Pengfei Li, Baojiang Chen, Peng Zhang and Zhijian Chen and many other fine graduate students in this department for their help and friendship during my graduate career.

This research was partially supported by Ontario Graduate Scholarship from the Ontario government and Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC), which allowed me to focus my energy and effort to this thesis research.

Finally, I wish to thank my husband and parents for their love and support, without which

I could not get this far. I also would like to thank my son, who was born during this thesis research. He helps me learn more about myself.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statisticians use models to approximate the relationship between a set of covariates Z and the variable of interest Y. The two main objectives of model building are to explain and to predict variable Y, using the covariates Z. Much of the statisticians' effort has been devoted to the estimation of model parameters and the construction of confidence intervals for them. However practitioners and scientists often desire a model that is reasonably easy to construct and interpret, and predicts well. Thus, the assessment of prediction performance is critical and has practical relevance. This is especially true for models with prediction as their primary objective. A good example is fraud detection in credit card transactions.

This thesis deals with the assessment of prediction performance of survival models. Survival models, also known as failure time models, provide statistical methods for the analysis and prediction of data when the variable of interest is the time to some event. The time variable Y is typically subject to censoring, which is the distinguishing and challenging feature of survival data. This type of data arises naturally from a wide class of settings, most notably in industry and biomedical sciences. In biomedical research, an important objective of multivariate survival time modeling is risk and survival time prediction for an individual with given covariates Z. For example, at the time of diagnosis, cancer patients are often informed of their expected survival time or survival probabilities at some specific future time. Clinicians themselves often use the predicted risk of a patient to classify him/her into some risk group and make treatment decisions. Thus, it is of great interest to assess the prediction accuracy of survival models and risk group classification.

The evaluation of prediction performance for survival models has been studied by various authors, e.g. Korn and Simon 1990; van Houwelingen and le Cessie 1990; Henderson 1995; Graf et al. 1999; Schemper and Henderson 2000; Heagerty et al. 2000; Rosthøj and Keiding 2004; Heagerty and Zheng 2005; Gerds and Schumacher 2006 & 2007; and Uno et al. 2007. Different frameworks and approaches have been proposed, which are reviewed in the second chapter. However, model misspecification and variable selection, which are common in practice, have not been fully addressed. In addition, further development for the treatment of censoring is needed along with methods for obtaining confidence intervals for prediction error. These are the problems we study in this thesis. In the next section, we first introduce two types of predictors, and then discuss the loss function approach for the evaluation of predictors.

1.1 Statistical Prediction

The standard prediction problem arises in the following scenario. We want to make prediction for the future random variable Y given covariates $Z = z$. To do so, a sample $D = \{(y_i, z_i), i = 1, \ldots, n\}$ is randomly selected from the population. Data D is often called the training data. The data analyst chooses a prediction procedure M and applies it to the training data to give a predictor for Y. Note that the prediction procedure may include model specification, parameter estimation, variable and model selection, and possibly tuning parameter selection, etc. We want to know how accurately this predictor predicts.

There are two types of predictors, point predictors and probabilistic predictors. A point predictor $\hat{Y}(Z) = G(Z; \hat{\theta})$ specifies a value for Y, where $\hat{\theta} = \hat{\theta}(D)$ is estimated from the training data D and $G(Z)$ is a function of covariates Z. A probabilistic predictor $\hat{F}_p(y|z) = \widehat{Pr}(Y \leq y \mid z)$ gives probabilities or prediction intervals for Y. In the aforementioned example, for a recently diagnosed cancer patient, a point predictor would be the predicted survival time, and a probabilistic predictor could be the predicted one year survival probability $\widehat{Pr}(Y > 1 \text{ year} \mid Z = z) = 1 - \hat{F}_p(1|z)$. Note that both types of predictors could be data or expert opinion based. To assess the accuracy of these two types of predictors, a common approach is to use a loss function.

It has been noted by some authors (e.g. see Graf et al. 1999) that probabilistic predictors are more useful than point predictors in many survival settings. Nevertheless, there are reasons to consider point prediction. Among them are the desire to classify individuals according to their predicted survival time in some settings; for example, the allocation of donor organs for transplantation is often based in part on predicted survival times for potential recipients. A second reason for interest in point predictors is the close relationship between their performance and measures of explained variation due to a set of covariates, which is reviewed in section 1.1.2. Our main focus in this thesis is point prediction, but probabilistic prediction is considered in chapter 6.

Some authors studying survival model prediction do not distinguish between these two types of predictors. In particular, the probabilistic predictor $\hat{S}_p(t|z) = \widehat{Pr}(Y > t \mid z)$ is considered by a number of authors (e.g. Schemper and Henderson 2000), and is treated as a point predictor for the indicator variable $W_t = I(Y > t)$. W_t is termed the survival status at time t; it takes value one if the survival time is greater than t and zero otherwise. In this thesis we require that a proper point predictor have the same support as the variable being predicted, which implies that $\hat{W}_t$ should only take value 0 or 1, and should not be an arbitrary probability.

1.1.1 Prediction loss and prediction error

A loss function approach is often used to evaluate the accuracy of a point predictor $\hat{Y}$. Let $L(Y, \hat{Y})$ denote the loss incurred when $\hat{Y}$ is used to predict a random variable Y. A loss function $L(\cdot)$ usually satisfies the following conditions: it is bounded below by 0 and attains 0 when correct prediction is made, i.e. $\hat{Y} = Y$; and as the "distance" between Y and $\hat{Y}$ increases, the loss function is nondecreasing. Two commonly used loss functions for a continuous variable Y are

$$\text{Squared error loss: } L(Y, \hat{Y}) = (Y - \hat{Y})^2,$$
$$\text{Absolute loss: } L(Y, \hat{Y}) = |Y - \hat{Y}|.$$

For a categorical variable W, we often use

$$\text{0-1 loss: } L(W, \hat{W}) = I(W \neq \hat{W}) = |W - \hat{W}|,$$

where $I(\cdot)$ is the indicator function and $\hat{W}$ is also categorical. In specific settings, other suitable loss functions may be defined, which is the case for survival data. One major challenge for survival data is to identify appropriate loss functions and there have been many discussions in the literature (e.g. Korn and Simon 1990, Henderson 1995, Henderson et al. 2001, Rosthøj and Keiding 2004). We leave the details to chapter 2.

For a point predictor $\hat{Y}$, the prediction error is defined as the expected loss. We call the predictor that minimizes the prediction error the optimal predictor. It is easy to show that the optimal predictor for squared error loss is $\hat{Y} = E_Y(Y)$ and for absolute error loss is $\hat{Y} = \text{median}(Y)$. The optimal predictor for a binary variable W, under 0-1 loss, is $\hat{W} = I(Pr(W = 1) > 0.5)$. Here we consider the case where the true distribution is known and used.

In practice, we base a point predictor on training data D, as follows. Suppose the true joint distribution of (Y, Z) is $F_T(y|z)H_Z(z)$, where $F_T(y|z)$ denotes the true conditional distribution function of Y given Z, and $H_Z(z)$ denotes the marginal distribution of Z. Training data $D = \{(y_i, z_i), i = 1, \ldots, n\}$ is a random sample from the joint distribution (Y, Z). To model the data, the analyst applies a modeling procedure M, which specifies some family of models $F_\theta(y|z)$, indexed by parameter θ, for the approximation of $F_T(y|z)$. Model $F_\theta(y|z)$ can be semiparametric or nonparametric, though we consider mainly parametric model in this thesis. Let $G(Z;\theta)$ denote the optimal predictor for Y under $F_\theta(y|z)$; that is, $G(Z;\theta)$ minimizes the prediction error $E_{F_\theta}[L(Y, G(Z))]$ among all functions $G(Z)$. $G(Z;\theta)$ is a function of F_θ; for example, if squared error loss is used, $G(Z;\theta) = E_{F_\theta}(Y|Z) = \int_{-\infty}^{\infty} y dF_\theta(y|Z)$.

The modeling procedure M yields $F_{\hat{\theta}}$ based on D. Therefore, the optimal predictor is given by $\hat{Y}(Z) = G(Z;\hat{\theta})$. We emphasize that the procedure M includes parameter estimation, variable and model selection, and possibly tuning parameter selection etc., thus $\hat{\theta} = \hat{\theta}(D)$ is usually a complex function of the training data D.

We want to know how well $G(Z;\hat{\theta})$ predicts on an independent test data set $D_{test} = \{(y_j, z_j), j = 1, \ldots, m\}$, arising from the same population (Y, Z). The prediction loss $L(y_j, G(z_j;\hat{\theta}))$ measures the prediction accuracy of the predictor $G(Z;\hat{\theta})$ for a realized Y_j given $Z = z_j$. And the average loss $m^{-1}\sum_{j=1}^{m}[L(y_j, G(z_j;\hat{\theta}))]$ measures the prediction accuracy of $G(Z;\hat{\theta})$ for the particular test data set.

The set up of the prediction problem leads to the following definitions,

$$\pi_1(M; F_T, z, D) = E_Y[L(Y, \hat{Y}(Z)) \mid Z = z, D] = \int L(y, G(z; \hat{\theta}))dF_T(y|z), \tag{1.1}$$

$$\pi_2(M; F_T, z) = E_{Y,D}[L(Y, \hat{Y}(Z)) \mid Z = z] = E_D[\pi_1(M; F_T, z)], \tag{1.2}$$

$$\pi_3(M; F_T, H_Z) = E_{Y,D,Z}[L(Y, \hat{Y}(Z))] = E_Z[\pi_2(M; F_T, Z)], \tag{1.3}$$

where $Y \sim F_T$ for $Z = z$, and is independent of training data D and therefore $\hat{Y}(Z)$. Here $\pi_1(M; F_T, z)$ measures the prediction accuracy of the procedure M given the training data D and that $Z = z$. Taking the expectation of (1.1) with respect to D, i.e. allowing training data to vary, we obtain the expected loss (1.2) that measures the performance of the procedure M under F_T for $Z = z$. This expectation is typically complex since $\hat{\theta}(D)$ is a complex function of D, as we discussed previously. Therefore, $\pi_2(M; F_T, z)$ generally has to be evaluated numerically, even when F_T is known. Finally we take expectation of (1.2) over the distribution of Z to give (1.3). Since H_Z is typically unknown, we frequently estimate the prediction error for the empirical Z distribution $\tilde{H}_Z$, based on $(z_1, \ldots, z_n)$ in D, that is,

$$\pi_3(M; F_T, \tilde{H}_Z) = \frac{1}{n}\sum_{i=1}^{n} \pi_2(M; F_T, z_i). \tag{1.4}$$

Here we focus on the estimation of $\pi_3(M; F_T, H_Z)$, and often consider (1.4), since $\pi_3(M; F_T, H_Z)$ measures the average performance of a prediction procedure M, under the true distribution F_T and H_Z. For simplicity, from now on we denote $\pi_3(M; F_T, H_Z)$ by π or $\pi(M; F_T)$, suppressing H_Z.

1.1.2 Explained variation and prediction power

Marginal prediction error

The prediction error π of a procedure M is a positive number. It is often useful to compare π with the prediction error of a model $F_0(y)$ not using the covariates Z. This type of model is known as marginal model or null model. Let $\hat{Y}_0$ denote the optimal predictor of Y based on $F_0(y)$, that is, $\hat{Y}_0$ minimizes the marginal prediction error,

$$E_Y[L(Y, \hat{Y}_0)] \leq E_Y[L(Y, \hat{Y})]$$

for all $\hat{Y}$. Let π_0 denote the marginal prediction error $E_Y[L(Y, \hat{Y}_0)]$. The ratio π/π_0 indicates how much prediction error is reduced when the covariates Z and the prediction procedure M are used for predicting Y.

Typically, the optimal marginal point predictor $\hat{Y}_0$ is a simple function of $\{Y_i, i = 1, \ldots, n\}$ in D. For example, $\hat{y}^0 = \sum_{i=1}^n y_i/n$ if we use squared error loss, and $\hat{y}^0 = \text{median}(y_i),\ i = 1, \ldots, n$, for absolute error loss.

Prediction power

Many authors have discussed the concepts of predictive power and explained variation of survival models (e.g. Korn and Simon 1990; Korn and Simon 1991; Schemper and Stare 1996; Schemper and Henderson 2000), which involves a comparison of the prediction error for regression models and the marginal prediction error. A common measure of predictive power, proposed by Korn and Simon (1991), is defined as

$$U = 1 - \frac{\pi}{\pi_0}. \tag{1.5}$$

U takes values between zero and one. It is close to zero when the denominator and numerator are close. That is, the marginal prediction error π_0 is almost as small as the prediction error π. In our framework, it suggests that the modeling procedure M and covariates Z do not have much prediction power for Y. On the other hand, U close to one suggests that π is much smaller than π_0, indicating that the use of procedure M and covariates Z leads to much improved prediction of Y.

By definition, U depends on $F_T(y|z)$, $H_Z(z)$, and the parameters indexing them, as well as the estimated model $F_{\hat{\theta}}(y|z)$. For illustration, we assume that F_T belongs to the model family F_θ, that is, $F_T = F_{\theta_0}$ for some θ_0 and assume θ_0 is known. Then for squared error loss, $\hat{Y}(Z) = E_{F_{\theta_0}}(Y|Z)$ and $\hat{Y}_0 = E_{F_{\theta_0}, H_Z}(Y)$, so (1.5) becomes

$$U = 1 - \frac{E_Z[\text{var}(Y|Z)]}{\text{var}(Y)} = \frac{\text{var}(E_{F_{\theta_0}}(Y|Z))}{\text{var}(Y)}. \tag{1.6}$$

Thus, U measures the percentage of variation in Y that is explained by the true model F_{θ_0}, hence the name “explained variation” (Korn and Simon 1991). We use “prediction power” and “explained variation” interchangeably hereafter.

For a normal linear regression model $Y = \alpha + \beta^T Z + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$,

$$\begin{aligned}
\hat{Y}(Z) &= \alpha + \beta^T Z, \\
\hat{Y}^0 &= E(Y) = \alpha + \beta^T E(Z), \\
\text{var}(Y|Z) &= \sigma^2, \\
\text{var}(Y) &= E[\text{var}(Y|Z)] + \text{var}[E_{F_{\theta_0}}(Y|Z)] = \sigma^2 + \beta^T \Sigma_Z \beta,
\end{aligned}$$

where Σ_Z denotes the covariance matrix of Z. In this case, equation (1.6) can be expressed as

$$U = \frac{\beta^T \Sigma_Z \beta}{\sigma^2 + \beta^T \Sigma_Z \beta}. \tag{1.7}$$

We see that U depends on the parameters β and σ^2, which index the conditional distribution of Y given Z, as well as Σ_Z, which indexes the distribution of Z.

1.1.3 Prediction error and prediction power of misspecified model

The true conditional distribution function $F_T(y|z)$ is rarely known, neither its functional form nor the parameter indexing it. We choose some model family $F_\theta(y|Z)$ based on the available data to approximate $F_T(y|z)$. If the maximum likelihood method is used for estimating θ, White (1982) proves that under regularity conditions, $\hat{\theta} \to \theta^*$ in probability as sample size approaches infinity, and θ^* is the unique solution to the estimating equation of score function $E_{F_T}[S(\theta)] = 0$. This θ^* minimizes the Kullback-Leibler divergence $D(\theta)$ (White 1982). $D(\theta)$ is a measure of the distance from the true unknown density to the density determined by F_θ (Kullback and Leibler 1951). It is defined as

$$\begin{aligned} D(\theta; Z) &= E\left\{\log\left[\frac{f_T(y|Z)}{f_\theta(y|Z)}\right]\right\} = \int f_T(y|Z)\log\left[\frac{f_T(y|Z)}{f_\theta(y|Z)}\right] dy, && (1.8) \\ D(\theta) &= E_{H_Z}\{D(\theta; Z)\}, && (1.9) \end{aligned}$$

where $f_T(y|z) = dF_T(y|z)/dy$ denotes the true unknown density and $f_\theta(y|z) = dF_\theta(y|z)/dy$ denotes the density function indexed by parameter θ. $D(\theta)$ is the average of distance $D(\theta; Z)$ over the distribution of Z. The parameter θ^* that minimizes $D(\theta)$ is sometimes called the least false parameter and $F_{\theta^*}(y|z)$ is the best approximating model to $F_T(y|z)$ in the model family F_θ.

Let $\hat{Y}^*(Z) = G(Z; \theta^*)$ denote the optimal point predictor given by the best approximating model $F_{\theta^*}(y|z)$. The associated prediction error $\pi_{\theta^*} = E_{F_T,H_Z}[L(Y, \hat{Y}^*(Z))]$ measures the prediction error of the best approximating model $F_{\theta^*}(y|z)$ with respect to the true unknown joint distribution of (Y, Z).

The optimal predictor of marginal model, $\hat{Y}_0$, remains unchanged, since it is often a simple function of the marginal distribution of Y. Hence, the prediction power of the misspecified model $F_{\theta^*}(y|z)$ with respect to the true data generating mechanism $F_T(y|z)$ is

$$U_{\theta^*} = 1 - \frac{\pi_{\theta^*}}{\pi_0}. \tag{1.10}$$

It measures the proportion of the prediction error π_0 for the null model that is reduced by the misspecified model $F_{\theta^*}(y|z)$. The greater U_{θ^*} is, the better the predictor $G(Z;\theta^*)$ predicts. Note that $\pi_T \leq \pi_{\theta^*}$, where π_T denotes the prediction error of the true model. This inequality can be verified by the definition of the optimal predictor. Let $\mu(Z)$ denote the optimal predictor for $L(Y, \hat{Y}(Z))$ under the true model $F_T(y|z)$, then

$$\pi_T = E_{F_T,H_Z}[L(Y, \mu(Z))] \leq E_{F_T,H_Z}[L(Y, \hat{Y}(Z))]$$

for all $\hat{Y}(Z)$ which include $\hat{Y}^*(Z)$. Consequently, the prediction power of model $F_{\theta^*}(y|z)$ is less than or equal to the prediction power of the true model $F_T(y|z)$, i.e. $U_{\theta^*} \leq U_{F_T}$.

1.2 Estimation of Prediction Error and Prediction Power

We now discuss estimation of prediction error π. If "test" data is available for the same set of Z values, i.e., $D_{test} = \{(y_i', z_i), i = 1, \ldots, n\}$, where y_i' is a realization from $F_T(y|z_i)$, the prediction error π can be estimated via

$$\hat{\pi} = \mathrm{L}_{\mathrm{test}} = \frac{1}{n}\sum_{i=1}^{n} L(y_i', G(z_i; \hat{\theta})). \tag{1.11}$$

When we do not have test data, the following three methods are the main approaches for the estimation of prediction error (e.g. Korn and Simon 1991, Efron 2004, Rosthøj and Keiding 2004).

1.2.1 Estimators for prediction error

Model-based method

The first method is based on estimating $\pi_2(M; F_T, z)$ (1.2) and $\pi(M; F_T)$ (1.3 or 1.4) by using an estimator $\hat{F}_T$, from which D and D_{test} (given $z_1, \ldots, z_n$) are assumed to arise.

Suppose $\hat{F}_T$ is obtained from D, the model-based estimator $\hat{\pi}^m$ takes the form

$$\hat{\pi}^m = \pi(M; \hat{F}_T) = \frac{1}{n} \sum_{i=1}^{n} E_{\hat{F}_T} \int_{-\infty}^{\infty} L(y, G(z_i; \hat{\theta})) d\hat{F}_T(y|z_i), \tag{1.12}$$

note that the expectation with respect to $\hat{\theta} = \hat{\theta}(D)$ is taken using $\hat{F}_T$. Again since variable and model selection are used in determining $\hat{\theta}(D)$, (1.12) can not be simplified in general and has to be evaluated to a desired degree of approximation by simulation.

A crucial question for the model-based approach is what to use for $\hat{F}_T$. There are various proposals around and the final model $F_{\hat{\theta}}$ given by the prediction procedure M is often used. A number of authors have argued that a sensibly chosen and adequately checked model can be expected to perform well, based on their experiences (e.g. see the discussion of Efron, 2004 and Rosthøj and Keiding 2004). However, this approach can be problematic, which is discussed in sections 2.3 & 3.5 and illustrated with the simulation results in chapters 3 & 4.

The fact that $\hat{\pi}^m$ is a function of $\hat{F}_T$ makes it convenient under some settings, e.g. in the case of the survival model where censoring creates difficulties for the use of other estimators, or when the distribution of Z for the population we want to predict is different from the distribution of Z in the training data, as we will see later in sections 3.2 and 7.1.

Apparent loss based method

By definition, prediction error measures the accuracy of a point predictor given by some modeling procedure on new independent data. Such new data is often not available. Alternatively, we can use the penalty method approach that does not require a new data set. This approach uses the "apparent loss" evaluated on the training data,

$$\mathrm{AL} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, G(z_i;\hat{\theta})), \tag{1.13}$$

and adds a penalty term to it. In (1.13), $G(z_i;\hat{\theta})$ is the predicted value for y_i of the training data and y_i is also from the training data. On its own, AL tends to underestimate (1.4) because Y_i and $G(z_i;\hat{\theta})$ are not independent; they are positively correlated.

Let Ω denote the bias, the difference between prediction error and the expected apparent loss. That is,

$$\Omega \;=\; \pi - E_D(\mathrm{AL}). \tag{1.14}$$

Consider a normal linear model with fixed number of covariates and no model misspecification,

$$Y = \beta^T Z + \varepsilon, \quad \varepsilon \sim N(0,\ \sigma^2).$$

Then under squared error loss,

$$\begin{aligned} \pi &= E_{Y,D}(Y - \hat{\beta}^T Z)^2 \\ &= E_Y(Y - \beta^T Z)^2 + E_D(\hat{\beta}^T Z - \beta^T Z)^2 \\ &= (1 + \frac{p}{n})\sigma^2, \end{aligned}$$

where p is the dimension of Z and n is the sample size. The expected apparent loss, on the other hand, is

$$\begin{aligned} E_D(AL) &= E_D(Y - \hat{\beta}^T Z)^2 \\ &= (1 - \frac{p}{n})\sigma^2. \end{aligned}$$

It turns out that $\Omega = \pi - E_D(AL) = \frac{2p}{n}\sigma^2$, which can be estimated by plugging in $\hat{\sigma}^2$. The linear normal model above is the simplest case and we can get an exact expression for the penalty term. However, if variable selection is involved and/or another loss function is used, an exact expression for the penalty term is typically difficult to obtain and can not be estimated directly.

A number of authors have proposed methods for estimating Ω (e.g., Mallows 1973; Stein 1981; Efron 1983; Efron 1986; Efron and Tibshirani 1997; Ye 1998; Tibshirani and Knight 1999 and Efron 2004), and the estimator given by Efron (2004) is used here. It applies to several loss functions and a wide class of models, and is given by

$$\Omega = \frac{2}{n}\sum_{i=1}^{n} \mathrm{cov}(y_i, f(\hat{y}_i)), \tag{1.15}$$

where the functional form of $f(\cdot)$ depends on the loss function used. For the following commonly used loss functions, Efron (2004) showed

- Squared error loss: $f(\hat{y}) = \hat{y} - 1/2$,
- Binary 0-1 loss (for binary variable Y): $f(\hat{y}) = -1/2$ or $1/2$ as $\hat{\mu}_y = \widehat{Pr}(Y = 1)$ is less than or greater than $1/2$,
- Entropy loss (for binary variable Y): $f(\hat{y}) = \log(\hat{\mu}_y/(1 - \hat{\mu}_y))$.

Efron (2004) uses the parametric bootstrap to estimate the covariance between y_i and $f(\hat{y}_i)$.

The size of Ω depends on the modelling procedure, the number of parameters being estimated, sample size, the dispersion of F_T, as well as the loss function used, among other factors (see for example, Ye 1998, Tibshirani and Knight 1999, and Efron 2004). Generally speaking, the more data adaptive the modelling procedure is, the more the number of parameters being estimated, and the smaller the sample size is, the bigger the penalty term is.

A penalty adjusted apparent loss estimator of prediction error is then, by (1.14) and (1.15)

$$\hat{\pi}^A = \mathrm{AL} + \hat{\Omega} = \mathrm{AL} + \frac{2}{n}\sum_{i=1}^{n} \widehat{\mathrm{cov}}(y_i, f(\hat{y}_i)). \tag{1.16}$$

Cross-validation method

The third approach is to use some form of cross-validation (CV) or data-splitting. It is the most widely used technique for estimating prediction error.

V-fold cross-validation splits the n training individuals into V sets S_v of approximately equal sizes n_v $(v = 1, \ldots, V)$ and uses the estimate

$$\hat{\pi}^{cv} = \frac{1}{n}\sum_{v=1}^{V}\sum_{i \in S_v} L(y_i, G(z_i; \hat{\theta}_{(-v)})), \tag{1.17}$$

where $\hat{\theta}_{(-v)} = \hat{\theta}(D/S_v)$ is obtained by applying the modeling procedure M to the training data D with S_v omitted. When V equals the sample size n, it is called "leave-one-out" cross-validation and $\hat{y}_{(-i)} = G(z_i; \hat{\theta}_{(-i)})$ is the predictor of y_i based on data

$$D/(y_i, z_i) = \{(y_1, z_1), \ldots, (y_{i-1}, z_{i-1}), (y_{i+1}, z_{i+1}), \ldots, (y_n, z_n)\}$$

with (y_i, z_i) excluded from D. Depending on n and V, (1.17) tends to overestimate (1.4) somewhat, because for $i \in S_v$ the predictor is based on a training data set of size $n - n_v$, rather than n.

The cross-validation approach enjoys the following properties that popularize its usage:

- It is a nonparametric method where no modelling assumption is made, and therefore it is robust to model misspecification;
- The estimator $\hat{\pi}^{cv}$ tends to have low bias;
- It is easy to implement.

Many implementations of statistical methods, e.g. partial least squares, classification and regression trees, nearest neighbor methods etc., use the cross-validation estimates $\hat{\pi}^{cv}$ for model selection. However, Efron (2004) showed that $\hat{\pi}^{cv}$ tend to be highly variable when compared to the estimates based on apparent loss $\hat{\pi}^{A}$. This trade-off between robustness and efficiency is common to all estimation problems, and can be investigated in specific settings. In addition, the performance of cross-validation estimates depends somewhat on the number of folds V, as does the amount of computation needed.

Other popular estimators of prediction error include the 0.632+ bootstrap estimator, the Monte Carlo cross-validation estimator, and etc. (e.g. Efron 1983, Efron and Tibshirani 1997, Molinaro et al. 2005, Tian et al. 2007, Gerds and Schumacher 2007). These estimators are in many ways similar to $\hat{\pi}^{A}$ or $\hat{\pi}^{cv}$ and can be investigated similarly.

1.2.2 Estimators for prediction power

We have reviewed the three common approaches for prediction error estimation. The same approaches can be used for the estimation of marginal prediction error π_0 and hence giving estimators for prediction power U.

The model based estimator uses the estimated distribution $\hat{F}_T$, generating both D and D_{test}. It takes the form,

$$\hat{\pi}_0^m = \frac{1}{n}\sum_{i=1}^{n} E_{\hat{F}_T,\tilde{H}_Z}[L(Y_i, \hat{Y}_0)]$$

where $\hat{Y}_0 = E_{\hat{F}_T,\tilde{H}_Z}(Y)$. The corresponding estimator for explained variation is

$$\hat{U}^m = 1 - \frac{\hat{\pi}^m}{\hat{\pi}_0^m}. \tag{1.18}$$

The estimator $\hat{U}^m$ is called estimated explained variation by Rosthøj and Keiding (2004). It only depends on the observed values $(y_i,\ z_i)$ through $\hat{F}_T$ and the empirical distribution of covariates $\tilde{H}_Z$.

The second estimator of the marginal error is based on apparent loss,

$$\hat{\pi}_0^A = \mathrm{AL}_0 + \hat{\Omega}_0, \tag{1.19}$$

where $\mathrm{AL}_0 = \frac{1}{n}\sum_{i=1}^n L(y_i, \hat{y}_0)$ and $\hat{\Omega}_0 = \frac{2}{n}\sum_{i=1}^n \widehat{\mathrm{cov}}(y_i, f(\hat{y}_0))$. $\hat{y}_0$ is the marginal predictor based on the empirical distribution of Y in the training data D. The explained residual variation (Korn and Simon 1991) is based on the apparent loss estimators,

$$\hat{U} = 1 - \frac{\mathrm{AL}}{\mathrm{AL}_0} = 1 - \frac{\sum_{i=1}^n L(y_i, \hat{y}_i)}{\sum_{i=1}^n L(y_i, \hat{y}_0)} \tag{1.20}$$

It can be shown that under appropriate conditions, $\hat{U}$ is a consistent estimator of U_{θ^*} (1.10) (Rosthøj and Keiding 2004). But in finite samples, the apparent losses tend to underestimate prediction error as we showed previously. We modify the estimator (1.20) slightly, replacing AL and AL_0 with the adjusted apparent loss estimator, that is

$$\hat{U}^A = 1 - \frac{\hat{\pi}^A}{\hat{\pi}_0^A}, \tag{1.21}$$

where $\hat{\pi}^A$ is defined in (1.16) and $\hat{\pi}_0^A$ is defined in (1.19). $\hat{U}^A$ is our second estimator for prediction power.

The cross-validation estimators of marginal error and prediction power are given by

$$\hat{\pi}_0^{cv} = \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v} L(y_i, \hat{y}_{0(-v)}), \tag{1.22}$$

$$\hat{U}^{cv} = 1 - \frac{\hat{\pi}^{cv}}{\hat{\pi}_0^{cv}}. \tag{1.23}$$

$\hat{y}_{0(-v)}$ is the marginal predictor based on the empirical distribution of Y in the training data D with S_v omitted.

1.3 Prediction Error and Survival Analysis

In survival analysis, prediction error and prediction power can be similarly defined. But the distinguishing feature of survival data, censoring, poses a unique challenge for the estimation of π with the loss function approach. The actual loss of a censored observation is not known because the corresponding survival time is unobserved. As a result, it is impossible to estimate the prediction error with $\hat{\pi}^A$ (1.16) or $\hat{\pi}^{cv}$ (1.17) unless a device such as imputation or weighting is used. We will review the current approaches that deal with censoring in the second chapter, and extend an inverse weighting method to accommodate arbitrary loss functions. The model-based estimator $\hat{\pi}^m$ can still be obtained since it only involves the expected loss computed under the model $\hat{F}_T$, and not the actual loss (1.12). Furthermore, we investigate how these estimators perform under model misspecification and variable selection with simulation studies.

Another challenge for survival data is to identify appropriate loss functions. Quite a few loss functions have been discussed and studied (e.g. Korn and Simon 1990, Henderson 1995, Henderson et al. 2001, Rosthøj and Keiding 2004). Henderson (1995) proposed several features of a desirable loss function for survival data. Some of them are as follows, when Y is a survival time:

1. The loss $L(Y, \hat{Y})$ should be bounded above as $|Y - \hat{Y}|$ increases. Consider the following scenario: if a subject survived only 1 month, the loss associated with a prediction of 20 years should be little greater than the loss with a prediction of 15 years. Likewise, the loss associated with a prediction of 1 month in comparison with an observed survival time of 20 years should be a little greater than the loss with lifetime being 15 years.

2. The bounds on $L(Y, \hat{Y})$ are not necessarily the same for $Y > \hat{Y}$ and $Y < \hat{Y}$.

3. The loss $L(Y, \hat{Y})$ does not have to be symmetric in $Y - \hat{Y}$. We may need the flexibility of treating underestimation and overestimation of survival time differently.

1.4 Thesis Outline

In this chapter, we defined prediction loss, prediction error, the associated concept of prediction power, and the estimation of the above quantities for a prediction procedure M. The squared error loss and normal linear model are used for illustration of various concepts. In the later chapters, we develop methods for obtaining estimators and confidence intervals for measures of prediction performance in survival models. The same methodology can be applied to obtain measures of prediction performance for marginal model $F_0(y)$, which are often used as references. Our contributions are: (i) we consider different approaches to the estimation of performance measures, recognizing that model selection and misspecification are ever present factors; (ii) we extend the inverse probability of censoring weights (IPCW) approach of Gerds and Schumacher (2006) to deal with arbitrary loss functions; (iii) we recognize that point estimates of prediction error are often subject to considerable uncertainty and thus we provide confidence interval procedures; (iv) we consider both point and probabilistic predictors and make connections for performance measures of the two types of predictors.

The remaining chapters are organized as follows. In the second chapter, we study the estimation of prediction error, including confidence intervals, for survival models. We first define appropriate loss functions, then we extend the inverse probability of censoring weighting (IPCW) approach for the censored survival data. In addition, we develop confidence interval procedures. In chapters 3 and 4, model-based, apparent loss based, and cross-validation estimators are compared through simulation studies, taking into account variable selection and model misspecification. Two loss functions, the absolute error loss and binary 0-1 loss, are investigated in the third and fourth chapters, respectively. Our results indicate that the model based methods are susceptible to model misspecification, but the apparent loss based and the cross-validation methods are robust. In the fifth chapter, we apply the methods to two survival data sets and give point and confidence interval estimates for prediction error. In the sixth chapter, we study the performance measures of probabilistic predictors for survival models, and connect the performance measures for the two types of predictors. Finally in the last chapter, we discuss the estimation of π when the distribution of covariates changes, and other future research topics.

Chapter 2

Estimation of Prediction Error in Survival Models

2.1 Loss Functions for Survival Data

A number of authors have discussed estimation of prediction error for survival models using the loss function approach (e.g. Korn and Simon 1990, Henderson 1995, Graf et al. 1999, Henderson et al. 2001, Rosthøj and Keiding 2004). An important question is which loss function to use, and the answer varies depending on the variable of interest. In survival analysis, the survival time Y is the response variable. But depending on the objectives of data analysis, other variables may be defined. In particular, the binary survival status W_t is often of interest, where $W_t = I(Y > t)$ indicates whether or not an individual is alive at some pre-specified time t. Clearly, the appropriate loss functions are different for variables Y and W_t, and we discuss them separately in section 2.1.1 and 2.1.2.

2.1.1 Loss function for survival time Y

In section 1.1.1, we mentioned that squared error loss and absolute error loss are often used for continuous variables. Between the two, squared error loss has received a great deal of attention in the literature, largely due to its mathematical convenience. It is easily decomposed and is differentiable. In addition the variance is defined using squared error loss, which makes the interpretation of results based on squared error loss easy to relate to. However, the squared error loss may not be a good choice for survival time Y, since survival distributions commonly found in practice are asymmetric with long tails to the right, for which the squared error loss tends to put too much weight on the extremely long-term survivors. For this reason, we prefer to use the absolute error loss for survival time Y. There are other benefits for choosing absolute error loss. From a practical point of view, median survival time can be easily obtained from the survival models and is widely reported in the medical literature. Note that the median is the optimal predictor for the absolute error loss. Therefore, we choose absolute error loss $L(Y, \hat{Y}) = |Y - \hat{Y}_m|$ for our investigation, where $\hat{Y}_m$ stands for the predicted median of Y. The prediction error based on it gives the expected value of the absolute difference between the future and predicted responses.

In many applications, accurate prediction of Y is thought to be important for subjects that are expected to die soon but of less interest for the long-term survivors. It is often enough to know that they will live for a long time. To incorporate this consideration, Korn and Simon (1990) proposed a bounded loss function. Instead of the unbounded absolute error loss, we consider

$$L^{\tau}(Y, \hat{Y}) = |(Y \wedge \tau) - (\hat{Y}_m \wedge \tau)|, \tag{2.1}$$

where $Y \wedge \tau = \min(Y, \tau)$ and τ is a specific time of interest, e.g. it could be the maximum followup time. With the bounded loss function, a zero loss is incurred if both Y and $\hat{Y}_m$ exceed

τ. Of course, setting $\tau = \infty$ gives the ordinary loss function $L(Y, \hat{Y})$. The loss function $L^\tau(Y, \hat{Y})$ (2.1) is bounded between $[0, \tau]$. This is equivalent to limiting the prediction error calculation to the range of [0, τ], which is desirable in survival settings, as outlined in section 1.3. Other types of bounded loss functions have been studied by Henderson (1995).

In addition to the truncation of loss functions, transformations of Y may be considered. Consider an observed survival time of 9 years with a prediction of 10 years and a survival time of 1 year with a prediction of 2 years; the former should result in a smaller loss. The log transformation is a convenient choice to address this concern. By using $\log(Y)$ instead of Y, the prediction error is invariant to the units of time and it gives relative errors that are often of more interest than absolute errors on the original time scale. We consider both transformation of Y and the truncation of the loss, the final form of our loss function for Y being

$$L^\tau(Y, \hat{Y}) = |\log(Y \wedge \tau) - \log(\hat{Y}_m \wedge \tau)| \tag{2.2}$$

Predictors $\hat{Y}_m$ are based on a model family $F_\theta(y|z)$ that approximates $F_T(y|z)$. An optimal predictor $G(Z; \hat{\theta})$ is obtained from the model $F_{\hat{\theta}}(y|z)$ fitted to training data D. For absolute error loss on the original scale (2.1) or on the log scale (2.2), we take $\hat{Y}_m(z) = G(z; \hat{\theta}) = \text{median}(Y|z; \hat{\theta})$, or $\underset{y}{\text{argmin}}\{F_{\hat{\theta}}(y|z) > 0.5\}$. Note that since $\text{median}(\log(Y)) = \log(\text{median}(Y))$, $\log(\hat{Y}_m) = \log(G(Z; \hat{\theta}))$ is the optimal predictor for $\log(Y)$ based on $F_{\hat{\theta}}(y|z)$. This is another nice feature of the absolute error loss.

2.1.2 Loss function for survival status W_t

In many survival settings, the variable of interest is not necessarily the survival time Y. For example, it is often of interest to identify patients who are going to suffer a relapse or succumb to disease early, say within 6 months. This time period is sometimes of special interest because of government regulation. In the UK and US, to qualify for hospice care, a patient usually has to be terminally ill with a life expectancy of 6 months or less. In another setting, cancer patients are sometimes considered cured if they are alive and cancer free 5 years after treatment. For both cases, we are interested to know whether patients' survival status at some future time t can be successfully predicted with covariates.

A binary variable W_t, known as survival status (e.g. Schemper and Henderson 2000), is defined as $W_t = I(Y > t)$ for pre-specified t. One possible predictor is the estimated survival probability $S_{\hat{\theta}}(t|Z) = Pr(Y > t \mid Z; \hat{\theta}) = 1 - F_{\hat{\theta}}(t|Z)$, for which the squared error loss and absolute error loss have been investigated (Korn and Simon 1990, Graf et al. 1999, and Schemper and Henderson 2000). However, it is not a proper point predictor for W_t since it does not have the same support as W_t, as discussed in section 1.1. The predictor that does this is the predicted survival up to t, $\hat{W}_t = I(S_{\hat{\theta}}(t|Z) > 0.5)$ or equivalently $I(\hat{Y}_m > t)$. By restricting $\hat{W}_t$ to be 0 or 1, we emphasize the ability of the predictor to classify individuals correctly as to whether they will or will not survive beyond time t.

A 0-1 or "misclassification error" loss function $L(W_t, \hat{W}_t)$ is

$$L(W_t, \hat{W}_t) = I(W_t \neq \hat{W}_t) = |W_t - \hat{W}_t|, \tag{2.3}$$

which here is equivalent to squared error loss. Although it is very often impossible to accurately predict Y for most individuals (e.g. Henderson et al. 2001 and Henderson and Keiding 2005), it is sometimes possible to more accurately predict who will survive beyond some time t; see for example, Korn and Simon (1990), Schemper and Henderson (2000), and Rosthøj and Keiding (2004).

2.1.3 ROC curve for W_t

When binary W_t is of interest, we essentially have a two-class classification problem. In addition to the misclassification error, the receiver operating characteristic (ROC) curve has been used extensively as a performance measure for binary classifiers, especially in medical diagnostic settings (Pepe 2003).

Consider a simple case where a continuous variable X is the only independent variable and a higher X value is more indicative of a class 1 subject, who is of interest to identify. Using a threshold c, subject i is classified into class 1 if and only if $X_i > c$, i.e. $\hat{W}_i = I(X_i > c)$.

There are two types of misclassification error:

$$\begin{aligned} \hat{W} = 1, \text{ but } W = 0 \text{ (false positive)},\\ \hat{W} = 0, \text{ but } W = 1 \text{ (false negative)}. \end{aligned}$$

The conditional probabilities $Pr(\hat{W} = 1|W = 0)$ and $Pr(\hat{W} = 0|W = 1)$ are termed false positive rate (FPR) and false negative rate (FNR), respectively. The true positive rate (TPR) and true negative rate (TNR) are defined as $Pr(\hat{W} = 1|W = 1)$ and $Pr(\hat{W} = 0|W = 0)$, respectively. Denote the false positive and true positive rates at threshold c as FPR(c) and TPR(c); then the ROC curve plots TPR(c) against FPR(c) as the threshold c moves through the range of X. As c decreases from ∞ to $-\infty$, the points on the ROC curve go from (0,0) to (1,1), which allows us to see both false positive and true positive rates for any given c.

A typical ROC curve is shown in Figure 2.1. The closer the ROC curve is to the left and top borders of the unit square, the better is the variable X at discriminating the two classes. A perfect test allows a complete separation of the two classes. That is, for some threshold c we have TPR(c)=1 and FPR(c)=0. The resulting ROC curve is along the left and top borders of

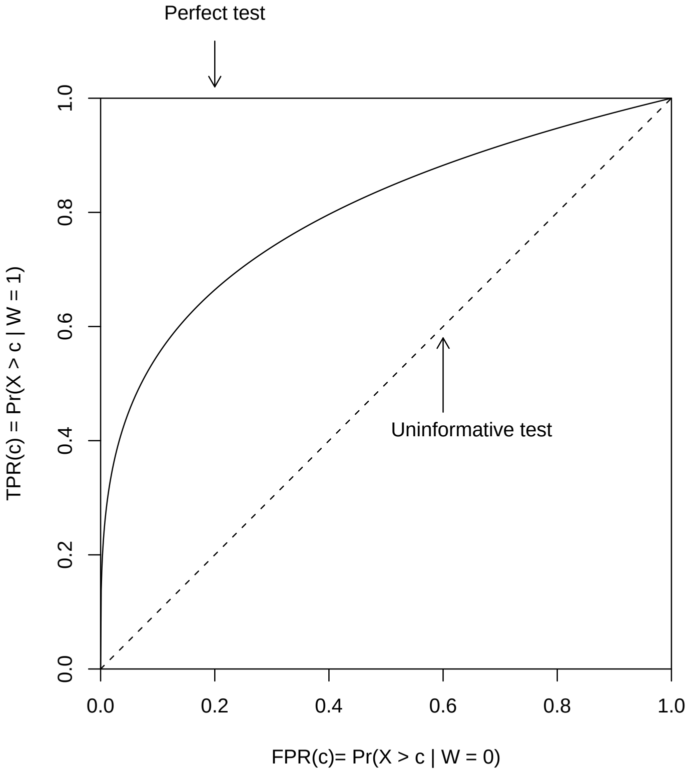

Figure 2.1: An ROC curve.

the unit square. On the other hand, an uninformative or useless test means that X is unrelated to W. The corresponding ROC curve is a straight line with unit slope. Note that the marginal model classifies all subjects into either class 0 or 1, which corresponds to either point (0,0) or (1,1) on the curve.

Put into the ROC context, the survival probability at time t, $S(t|z)$, takes the role of X. Classification or prediction rules are of the form

$$\hat{W}_t = I(S(t|z) > c), \tag{2.4}$$

where $c \in [0,\ 1]$. In section 1.1.1, we showed that the optimal predictor for 0-1 loss is rule (2.4) with $c = 0.5$. The 0-1 loss treats the false positive and false negative equally. When the two types of error incur different costs c_0 and c_1, it can be shown that the optimal predictor is (2.4) with $c = \frac{c_0}{c_0 + c_1}$. The ROC curve conveniently displays the entire set of possible c values in one graph. This is a good idea especially when we may not wish to focus completely on a single loss function.

A summary measure for ROC curves is the area under curve (AUC), which has a probability interpretation. That is,

$$\text{AUC} = Pr(S(t|z_i) > S(t|z_j) \mid T_i > t, T_j \leq t),$$

where the i^{th} individual is randomly selected from class 1 and the j^{th} individual is from class 0. However, this probability interpretation of AUC is not that useful because it is a conditional probability. Copas (1999) showed that the AUC is not a good measure for judging the usefulness of $S(t|z)$ when the two classes are highly unbalanced. This will be the case if we are interested in early failures which are only a small fraction of the population.

The ROC curve and the associated AUC statistic are often used to compare two or more classification models. But as Adams and Hand (1999) pointed out, only in the case that one classifier dominates another will the AUC be universally valid in a comparison of classification models. In reality, two competing classification methods may yield ROC curves that cross each other. If that happens, one classification model is better for some values of the cost ratio and the other model will be better for other values.

In the context of survival analysis, we use models to predict the survival probabilities $S_{\hat{\theta}}(t|z)$. ROC curve compares the ranking of these survival probabilities with the true survival status and shows how well the survival model identifies the class of interest. It may be more informative than the misclassification error under specific settings. For example, when t is small, only a small proportion of the population fails, and the marginal model classifies everyone into the survived category and results in a small misclassification error. The regression model may not improve the misclassification error much but may be able to correctly identify individuals with high risk of failure at t, and that could be visualized with the ROC curve.

2.2 Estimation of Prediction Error in the Presence of Censoring

Survival data is known for the censoring of survival time Y. Let Y_i denote the survival time and C_i denote a potential censoring time for the i^{th} individual. The data observed on n independent individuals is $D = \{(T_i, \delta_i, Z_i),\ i = 1, \ldots, n\}$, where $T_i = \min(Y_i, C_i)$ is the right censored survival time and $\delta_i = I(Y_i \leq C_i)$ is the censoring indicator (e.g. Lawless 2003). When $\delta_i = 1$, Y_i is observed. When $\delta_i = 0$, Y_i is censored, all we know is that the individual is still alive at C_i.

Consider the estimator (1.11), which is of the form

$$\hat{\pi} = \frac{1}{m} \sum_{j=1}^{m} L(y_j', \hat{y}_j), \tag{2.5}$$

where $\hat{y}_j = G(z_j; \hat{\theta})$ and the (y_j', z_j) are independent realizations from $F_T(y|z)H_Z(z)$. However, we can not use (2.5) if some of the Y_j are censored. This is a type of missing data problem. There are two general approaches to missing data problems: weighting and imputation. Both methods have been explored in survival settings, for example, by Graf et al. (1999), Schemper and Henderson (2000), Gerds and Schumacher (2006), and Uno et al. (2007).

2.2.1 Inverse probability of censoring weighting

Inverse probability weighting for censoring was introduced by Robins and Rotnitzky (1992) and used by Graf et al. (1999) to deal with censored losses in prediction error estimation. Uno et al. (2007) used the same weighting method for a different loss function and a different class of survival models. Both papers assume random censoring where C is completely independent of Y and Z, and propose estimators for the survival status W_t. Their estimators can be generalized to the survival time Y, which is given below for illustration of their method.

The inverse probability weighting method assign weights α_j^{-1} for the j^{th} individual whose survival time y_j is observed, where $\alpha_j = Pr(C \geq y_j) = S^c(y_j)$ is the probability that the j^{th} individual survived to time y_j without being censored. By definition, the weights $\alpha^{-1} \geq 1$ for uncensored individuals. Censored individuals, on the other hand, receive weights 0. Thus, they do not contribute to the estimator of prediction error directly. Their contributions are made through the weights, which are determined by the observed censoring times. The weighted estimator is of the form

$$\hat{\pi}^w = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\delta_j}{\alpha_j}L(y'_j, \hat{y}_j)\right), \tag{2.6}$$

where δ_j is the censoring indicator for the j^{th} individual, $\alpha_j = S^c(y'_j)$ is the marginal censoring time distribution for C, and y'_j and $\hat{y}_j$ are independent. Given that C is completely independent of Y and Z, $E_C(\hat{\pi}^w)$ equals $\hat{\pi}$ in (2.5). Under suitable conditions, Rosthøj and Keiding (2004) proved that $\hat{\pi}^w$ is a consistent estimator of π for (1.3) or (1.4).

The assumption of random censoring is quite restrictive and often not met in practice. Normally we only want to assume independent censoring, i.e., censoring times C are independent of survival times Y given Z. Gerds and Schumacher (2006) propose an estimator that can handle independent censoring for the loss $L(W_t, \hat{W}_t) = (W_t - \hat{W}_t)^2$.

Let

$$S^c(c|z) = Pr(C > c \mid Z = z) \tag{2.7}$$

denote the conditional survivor function of the censoring variable C given Z. The IPCW approach replaces α by an estimate of $S^c(c|z)$ in (2.6) and similar expressions,

$$\hat{\pi}^w = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\delta_j}{\hat{S}^c(y_j|z_j)}L(y_j, \hat{y}_j)\right). \tag{2.8}$$

Note that when we model the censoring distribution $S^c(c|z)$, it may be misspecified, in which case the estimator $\hat{S}^c(c|z)$ can be biased. This issue is briefly discussed in section 7.2.4. For now, we assume that $S^c(c|z)$ can be consistently estimated.

We now generalize the Gerds & Schumacher's IPCW estimator (2.8) to arbitrary loss functions, using the general approach of van der Laan et al. (2002). Our approach is similar to Gerds & Schumacher's in the sense that we also use the IPCW, but our method is more general and (2.8) can be viewed as a special case of our estimator.

Define the binary variable

$$\Delta_j = I\{L(Y_j, G(Z_j; \hat{\theta})) \text{ is known}\}, \tag{2.9}$$

and note that Δ_j depends on $Y_j,\ Z_j,\ C_j$ and training data D. Let

$$\alpha_j = Pr(\Delta_j = 1 \mid Y_j, Z_j, D), \tag{2.10}$$

and note further that given $Y_j,\ Z_j$ and D, the random variable Δ_j is a function of censoring time C_j only. Therefore α_j is determined by the conditional censoring distribution $S^c(c|z)$. For example, if $L(Y_j, \hat{Y}_j) = (Y_j - \hat{Y}_j)^2$ then $\Delta_j = \delta_j = I(C_j \geq Y_j)$ and $\alpha_j = S^c(Y_j|Z_j)$. If $L(Y_j, \hat{Y}_j)$ is the bounded loss function given by (2.1) or (2.2), then

$$\Delta_j = I(Y_j > \tau, C_j > \tau) + I(Y_j \leq \tau, Y_j \leq C_j)$$

and

$$\alpha_j = I(Y_j > \tau)S^c(\tau|Z_j) + I(Y_j \leq \tau)S^c(Y_j|Z_j).$$

Similarly, for $L(W_{t,j}, \hat{W}_{t,j})$,

$$\begin{aligned} \Delta_j &= I(W_{t,j} = 1) + I(W_{t,j} = 0, \delta_i = 1) \\ &= I(Y_j > t, C_j > t) + I(Y_j \leq t, Y_j \leq C_j) \qquad (2.11) \\ \alpha_j &= I(Y_j > t)S^c(t|Z_j) + I(Y_j \leq t)S^c(Y_j|Z_j). \qquad (2.12) \end{aligned}$$

Our IPCW estimator, for $\alpha_j > 0$ and arbitrary loss function $L(Y, \hat{Y})$ is given by

$$\hat{\pi}^w(M; F_T) = \frac{1}{m}\sum_{j=1}^{m} \frac{\Delta_j}{\hat{\alpha}_j} L(y_j, \hat{y}_j). \qquad (2.13)$$

where $\hat{\alpha}_j$ is an estimate of $S^c(\cdot|z)$ at a suitable time. The motivation for (2.13) is

$$E\left[\frac{1}{m}\sum_{j=1}^{m} \frac{\Delta_j}{\alpha_j} L(Y_j, \hat{Y}_j)\right] = E\left[\frac{1}{m}\sum_{j=1}^{m} L(Y_j, \hat{Y}_j)\right],$$

where the expectation on the left is with respect to $Y,\ Z,\ C$ and D, and the one on the right is with respect to $Y,\ Z$ and D. The above equation holds because

$$E_{Y,Z,C,D}\left[\frac{1}{m}\sum_{j=1}^{m} \frac{\Delta_j}{\alpha_j} L(Y_j, \hat{Y}_j)\right] = E_{Y,Z,D}\left[E_{C|Y,Z,G_{\hat{\theta}}}[\frac{1}{m}\sum_{j=1}^{m} \frac{\Delta_j}{\alpha_j} L(Y_j, \hat{Y}_j)]\right],$$

where $\alpha_j = E(\Delta_j|Y_j, Z_j, D)$ by the definition of α_j (2.10). Note that Y_j has to be a time for which $S^c(Y_j|z_j) > 0$ holds almost surely.

IPCW estimator of $\hat{\pi}^A$ and $\hat{\pi}^{cv}$

We now apply (2.13) to the cases of adjusted apparent loss estimator (1.16) and CV estimator (1.17) for survival time Y, and obtain

$$\hat{\pi}^A = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{\hat{\alpha}_i} L(y_i, \hat{y}_i) + \hat{\Omega} \qquad (2.14)$$

and

$$\hat{\pi}^{cv} = \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v}\frac{\Delta_i}{\hat{\alpha}_i}L(y_i, \hat{y}_{i(-v)}). \tag{2.15}$$

The consistency of the IPCW estimators (2.14) and (2.15) depends on whether $\hat{S}^c(c|z)$ is a consistent estimator of the conditional censoring distribution $Pr(C > c \mid z)$. Similar IPCW estimators for prediction error of marginal model can be obtained.

ROC curve estimation using IPCW approach

In section 2.1.3, we discussed using the ROC curve to evaluate the performance of a classification rule for survival status W_t. Define $\hat{W}_t = I(S(t|z) > c)$ as in (2.4), the true positive and false positive rates at c are then given by

$$\text{TPR}(c) = Pr(\hat{W}_t = 1 \mid W_t = 1) = \frac{Pr(S(t|z) > c, Y > t)}{Pr(Y > t)}, \tag{2.16}$$

$$\text{FPR}(c) = Pr(\hat{W}_t = 1 \mid W_t = 0) = \frac{Pr(S(t|z) > c, Y \leq t)}{Pr(Y \leq t)}, \tag{2.17}$$

for $c \in (0,1)$. When some of the W_t are censored, the ROC curve can not be estimated directly. Heagerty et al. (2000) discussed two estimation methods for (2.16) and (2.17). The first method uses the Bayes theorem and the Kaplan-Meier estimator to give $\widehat{Pr}(Y > t)$ and $\widehat{Pr}(S_{\hat{\theta}}(t|z) > c, Y > t)$. However, it is shown that the ROC curve estimated by this method may not be monotonic. The second method uses a nearest neighbor kernel method for the estimation of the bivariate function, $Pr(S_{\hat{\theta}}(t|z) > c, Y > t)$ (Akritas 1994). It can ensure the monotonicity of the estimated ROC curve and has been used by several authors (e.g. Li and Gui 2004 and Guo et al. 2006).

Here we show that the (2.16) and (2.17) can be estimated using the IPCW approach and the monotonicity of the resulting ROC curve is also guaranteed.

Consider the numerator of (2.16) $Pr(S(t|z) > c, Y > t)$, which normally can be estimated by $n^{-1}\sum_{i=1}^{n} I(Y_i > t, S_{\hat{\theta}}(t|z_i) > c)$. When censoring is present, we use instead

$$\hat{Pr}(S(t|z) > c, Y > t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{\alpha}_i} I(Y_i > t, S_{\hat{\theta}}(t|z_i) > c), \tag{2.18}$$

for $c \in (0, 1)$ and where Δ_i (2.11) and α_i (2.12) are defined for $L(W_t, \hat{W}_t)$. Other probabilities such as $Pr(Y > t)$, $Pr(S(t|z) > c, Y \leq t)$ and $Pr(Y \leq t)$ can be estimated the same way.

The monotonicity is guaranteed by noting that as c decreases, the right hand side of (2.18), $n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{\alpha}_i} I(Y_i > t, S_{\hat{\theta}}(t|z_i) > c)$, is nondecreasing, because the weights $\Delta_i/\hat{\alpha}_i$ are non-negative.

2.2.2 Imputation

Imputation is also used for the estimation of prediction error in survival data. Schemper and Henderson (2000) consider the survival status W_t and impute the censored loss with its expected value, which is determined by the estimated regression model, conditional on the censoring time C. Their prediction error estimator is given by averaging the available losses and the expected conditional losses, that is,

$$\hat{\pi}_t^{sh} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\Big(\Delta_i L(w_{i,t}, \hat{w}_{i,t}) + (1-\Delta_i)E_{\hat{\theta}}[L(W_{i,t}, \hat{W}_{i,t})|Y > c_i, Z = z_i]\Big).$$

Note that this estimator is a sum of both model based and non-model based terms.

Robins and his coworkers (e.g. Robins and Rotnitzky 1992, van der Laan et al. 2002) have shown that the weighting method has a number of advantages over the imputation method in dealing with missing data problems. Our simulation results in chapter 3 and 4 also suggest that the model-based terms could be seriously biased when the model is misspecified.

2.3 Probability Limits of Prediction Error Estimators

It is useful to consider what an estimator of prediction error converges to in probability when the size n of the training data sample becomes arbitrarily large. To do this, we assume that a well-defined prediction procedure M produces an estimate $\hat{\theta}_n(D)$ and a corresponding model $F_{\hat{\theta}_n}$ that gives predictor $G(Z; \hat{\theta}_n)$. This procedure would include, for example, the specification of model family F_θ and rules for the selection of covariates or tuning parameters. In letting n become large, we assume that covariate values $z_1, \dots, z_n$ in D are generated independently from a distribution H_Z and the corresponding y_i are generated independently from $F_T(y|z_i),\ i = 1, \dots, n$.

Authors such as Rosthøj and Keiding (2004), Gerds and Schumacher (2006), and Tian et al. (2007) provide certain analytical results. All the authors assume that estimates $\hat{\theta}$ are of fixed dimension and, in some cases, that the true distribution F_T is a member of the parametric family F_θ used to obtain the predictor. For differentiable loss functions and $F_T \in F_\theta$, Rosthøj and Keiding (2004) showed the weighted estimator $\hat{\pi}^w$ (2.6) by Graf et al. (1999) and the estimator $\hat{\pi}^{sh}$ by Schemper and Henderson (2000) are consistent estimators of prediction error (1.3). Under the misspecified survival model, i.e. F_T does not belong to F_θ, the estimator of Graf et al. is still consistent while the estimator of Schemper and Henderson is not. Tian et al. (2007) studied the absolute error loss and proved that the prediction error can be consistently estimated by apparent loss and V-fold cross-validation loss, under the condition that the regression coefficients converge to a limit and other regularity conditions. In Gerds and Schumacher (2006), the uniform consistency of the IPCW estimator (2.8) is established for a specific loss function $L(W_t, S_{\hat{\theta}}(t|z)) = (W_t - S_{\hat{\theta}}(t|z))^2$, under the assumption that the censoring distribution is correctly specified.

A rigorous development for the consistency of our IPCW estimator (2.13) would require a careful specification of the prediction rules. Here we provide a heuristic discussion only. This can, however, be checked by simulation and compared with finite sample properties. As in Rosthøj and Keiding (2004), we assume conditions on the training data and family F_θ of models so that $\hat{\theta}_n$ converges in probability to a limit θ^* as $n \to \infty$. The model-based estimator (1.12) will then, under suitable conditions like those assumed by Rosthøj and Keiding, converge to $\pi(M; F_{\theta^*})$. If $F_{\theta^*} = F_T$ then (1.12) estimates the "true" prediction error of the procedure and is a consistent estimator of $\pi(M; F_T)$. Of course, no model is true in practice and one hopes that a sensible procedure will give a well-specified model which produces a prediction error estimate that is not too biased.

The AL estimator (1.13) and (1.16) and CV estimator (1.17) do converge in probability to $\pi(M; F_T)$, the true prediction error for the procedure in question. Once again, this can be shown under conditions similar to those in Rosthøj and Keiding (2004). Because of censoring, we can not use (1.16) and (1.17), but consider instead the IPCW versions (2.14) and (2.15). To establish that they converge in probability to $\pi(M; F_T)$, we require a consistent estimator $\hat{S}^c(c|z)$ on which to base the weight $\hat{\alpha}_i$ in (2.14) and (2.15). Under suitable conditions we can then show that $n^{-1} \sum_{i=1}^{n} \xi_i L(Y_i, G(Z_i; \hat{\theta}_n))$ with ξ_i equal to each of 1, Δ_i/α_i and $\Delta_i/\hat{\alpha}_i$, all converge to the same probability limit, which is $\pi(M; F_T)$. Misspecification of $S^c(c|z)$ is discussed briefly in section 7.2.4.

In practice, there is often a subjective element to model specification and the selection of a predictor and so estimates of prediction error must be interpreted cautiously. In chapters 3 and 4, we carry out simulation studies with a fixed rule for variable selection and are able to provide support for the limits stated above.

2.4 Construction of Confidence Intervals

Estimates of prediction error are usually subject to considerable uncertainty. For example, consider the case of squared error loss, where the optimal predictor is $\mu(Z) = E_{F_T}(Y|Z)$. For a predictor $G(Z;\hat{\theta}) = \hat{\mu}(Z)$ given by procedure M, we have

$$\begin{aligned}\pi(M;F_T) &= E\{[Y - \hat{\mu}(Z)]^2\} \\ &= E_Z\{\text{var}(Y|Z) + \text{var}[\hat{\mu}(Z)] + [\mu^*(Z) - \mu(Z)]^2\},\end{aligned}$$

where $\mu^*(Z) = E_{F_T}[\hat{\mu}(Z)]$. To estimate π we have to estimate variances, and it is well-known that sample sizes must be reasonably large to do this precisely. Other loss functions are also based on measures of variation, and a similar situation holds. Nonetheless, there has been little discussion of confidence interval estimation, with Uno et al. (2007) being a recent exception. Rosthøj and Keiding (2004) acknowledged that they were unable to approximate the variance of the prediction error estimators well using the analytical approach; and there was no variable selection in their modelling procedure. Uno et al. (2007) also noted the difficulty of obtaining variance estimates, and they used a perturbation-resampling procedure to give an approximation to the variance estimate; their results were based on fixed models as well.

We propose to use bootstrap procedures to give confidence intervals for prediction error (1.3) and (1.4). The distributions of the prediction error estimators $\hat{\pi}^m$ (1.12), $\hat{\pi}^A$ (2.14) and $\hat{\pi}^{cv}$ (2.15) depend on the sampling variability of the training data D, and our confidence interval procedures are based on the generation of B pseudo training samples D_b^* $(b = 1, \ldots, B)$ by using either parametric or nonparametric bootstrap sampling.

Assume random censoring and suppose the parametric bootstrap is used for obtaining D_b^*. We use the following procedure to approximate the distribution of $\hat{\pi}^m$ (1.12), which is used later for the construction of confidence intervals for π.

1. Apply the prediction procedure M to D, giving the estimated model $F_{\hat{\theta}}(y|z)$; obtain the Kaplan-Meier estimate $\hat{S}^c$ for the survivor function of C.

2. Generate the survival times $Y_b^* = \{y_1^*, \ldots, y_n^*\}$ from $F_{\hat{\theta}}$ for $\{z_1, \ldots, z_n\}$, and the censoring times $C_b^* = \{c_1^*, \ldots, c_n^*\}$ from $\hat{S}^c$. Let $t_i^* = \min(y_i^*, c_i^*)$, which then gives $D_b^* = \{(t_i^*, \delta_i^*, z_i),\ i = 1, \ldots, n\}$.

3. Apply the same model selection procedure to D_b^*, which gives an estimated model $F_{\hat{\theta}_b}$. Estimate the survivor function $\hat{S}_b^c$ from D_b^* with the Kaplan-Meier method.

4. Repeat step 2, replacing $F_{\hat{\theta}}$ and $\hat{S}^c$ with $F_{\hat{\theta}_b}$ and $\hat{S}_b^c$ to produce D_{bk}^*; in addition, generate an independent set of survival times $Y_{bk}^* = \{y_i^{*bk},\ i = 1, \ldots, n\}$ from $F_{\hat{\theta}_b}$ for $\{z_1, \ldots, z_n\}$. The set Y_{bk}^* is used for assessing the average prediction loss in step 5.

5. Apply the prediction procedure to D_{bk}^*, which gives predictor $\hat{y}_i^{*bk} = G(z_i; \hat{\theta}_{*bk})$. Suppose the loss function is $L^\tau(Y_i, \hat{Y}_i)$, then the average prediction loss is

$$PL_{bk} = \frac{1}{n} \sum_{i=1}^{n} L^\tau(y_i^{*bk}, \hat{y}_i^{*bk}),$$

where $y_i^{*bk} \in Y_{bk}^*$.

6. Repeat steps 4 and 5 K times, obtaining a bootstrap estimate of prediction error (model-based) for D_b^*

$$\hat{\pi}_b^{*m} = \frac{1}{K} \sum_{k=1}^{K} PL_{bk}.$$

7. Repeat steps 2 through 6 B times, giving bootstrap estimates $\{\hat{\pi}_b^{*m},\ b = 1, \ldots, B\}$, which approximate the distribution of $\hat{\pi}^m$ and are used for the construction of confidence interval for π, as discussed later.

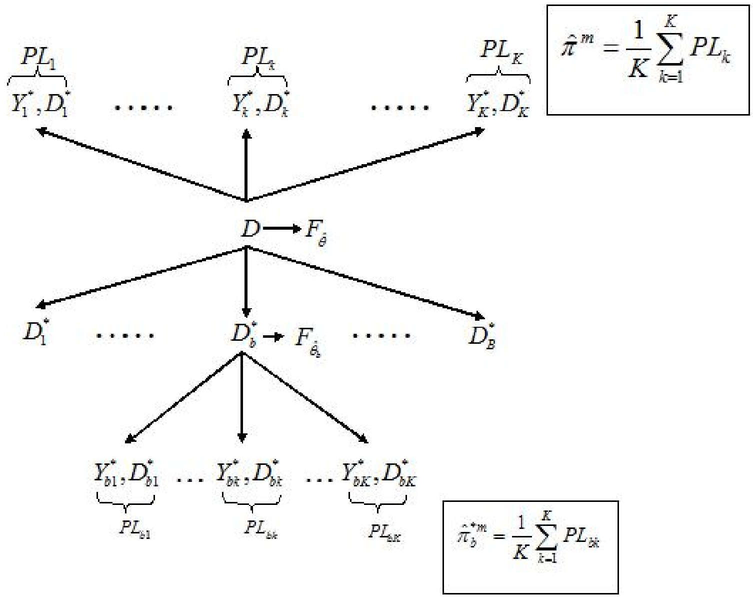

Figure 2.2: An illustration for the point and variance estimation of $\hat{\pi}^m$ using parametric bootstrap procedure.

The point estimate $\hat{\pi}^m$ is obtained following the above simulation procedure steps 3, 4, 5 and 6, with $F_{\hat{\theta}_b}$, $\hat{S}_b^c$, D_{bk}^* and Y_{bk}^* replaced with $F_{\hat{\theta}}$, $\hat{S}^c$, D_k^* and Y_k^*, respectively; see section 3.2 for a more detailed description. Figure 2.2 gives a graphical representation of the procedure above.

In settings where we want to consider independent censoring, we assume a model family $S_\theta^c(c|z)$ for survivor function $Pr(C > c \mid Z = z)$ for censoring variable C. Then instead of the Kaplan-Meier estimates $\hat{S}^c$ and $\hat{S}_b^c$, $S^c(c|z;\hat{\theta})$ fitted to D and $S_b^c(c|z;\hat{\theta}_b)$ fitted to D_b^* are used to generate the censoring times in steps 2 and 4, respectively. When nonparametric bootstrap is used instead of parametric bootstrap, D_b^* in step 2 consists of selecting n items from the original training data $D = \{(y_i, z_i), i = 1, \ldots, n\}$ with replacement. Note that we do not need to generate censoring times for D_b^* when nonparametric bootstrap is used.

There are two levels of bootstraps in Figure 2.2 for getting the bootstrap estimates $\hat{\pi}_b^{*m}$ $(b = 1, \ldots, B)$. In the first level, D_b^* could be obtained by either parametric or nonparametric bootstrap, but the second level has to be a parametric bootstrap in order to produce independent D_{bk}^* and Y_{bk}^*. Note that to obtain bootstrap estimates $\hat{\pi}_b^{*A}$ or $\hat{\pi}_b^{*cv}$ for $\hat{\pi}^A$ or $\hat{\pi}^{cv}$, only the first level of bootstrap is needed, which could be either parametric or nonparametric. This is because these two estimators can be obtained with only the pseudo training data D_b^*.

In the present setting, analytic variance estimates for $\hat{\pi}$ are not available, and for computational reasons we consider simple normal approximations, with the B bootstrap estimates $\hat{\pi}_b^*$ used to estimate var$(\hat{\pi})$ as

$$\widehat{\mathrm{var}}(\hat{\pi}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\pi}_b^* - \bar{\hat{\pi}}^*)^2, \quad \hat{\mathrm{sd}}(\hat{\pi}) = \widehat{\mathrm{var}}(\hat{\pi})^{1/2}$$

where $\bar{\hat{\pi}}^* = \sum_{b=1}^{B} \hat{\pi}_b^*/B$. We then treat $(\hat{\pi} - \pi)/\hat{\mathrm{sd}}(\hat{\pi})$ as a pivotal quantity that has a standard normal distribution, producing $1 - \alpha$ confidence intervals

$$\hat{\pi} \pm \Phi^{-1}(\alpha/2)\hat{\mathrm{sd}}(\hat{\pi}),$$

where $\Phi(\cdot)$ denote the cumulative density function of the standard normal distribution and $\Phi^{-1}(\alpha/2)$ gives the $\alpha/2$ quantile for the standard normal distribution.

We have also investigated bias corrections, replacing $\hat{\pi}$ with $2\hat{\pi} - \bar{\hat{\pi}}^*$. This is because the bias $E(\hat{\pi}) - \pi$ can be estimated by $\bar{\hat{\pi}}^* - \hat{\pi}$ (Efron and Tibshirani 1993, section 10.2). Another idea we experimented with is the alternative use of $\psi = \log \pi$ for the normal approximations, with confidence intervals based on $(\hat{\psi} - \psi)/\hat{\mathrm{sd}}(\hat{\psi})$. The variance estimate for $\hat{\psi}$ is obtained from the bootstrap estimates $\hat{\psi}_b^* = \log \hat{\pi}_b^*$. This may be a sensible choice since $\hat{\pi}$ is always positive and its distribution likely has a longer right tail. To see this, consider the χ^2 distribution, which has a close connection with $\hat{\pi}^A$ in the case of squared error loss under normal linear models. A log transformation may help normalize the distribution of $\hat{\pi}$, especially when sample size n is not very large. We have also considered the percentile method of obtaining confidence intervals (Efron and Tibshirani 1993, Section 13.3). Letting $\hat{\pi}_B^{*(\alpha)}$ be the $100 \cdot \alpha$th empirical percentile of the $\hat{\pi}_b^*$ values, the approximated $1-\alpha$ confidence interval is

$$[\hat{\pi}_B^{*(\alpha/2)},\ \hat{\pi}_B^{*(1-\alpha/2)}]. \tag{2.19}$$

Finally, we considered the "basic" percentile method (Davison and Hinkley 1997, Section 5.3), which is given by $[2\hat{\pi} - \hat{\pi}_B^{*(1-\alpha/2)}, 2\hat{\pi} - \hat{\pi}_B^{*(\alpha/2)}]$.

Procedures for obtaining model-based, AL and CV estimators (1.12), (2.14) and (2.15) and confidence intervals based on them are considered for the cases of absolute error and binary losses in chapters 3 and 4. Because the model-based estimators require simulation, confidence intervals based on them require two levels of simulation, as illustrated in Figure 2.2. Consequently, we typically use $B = 100$ bootstrap samples for confidence interval estimation, whereas with the AL and CV estimators a larger number such as 500 is computationally feasible.

Chapter 3

Estimation of Prediction Error with Absolute Error Loss

3.1 Absolute Error Loss for log Survival Time

In this chapter, we consider the bounded absolute error loss for $\log(Y)$ (2.2) defined in section 2.1.1,

$$L^{\tau}(Y, \hat{Y}) = |\log(Y \wedge \tau) - \log(\hat{Y}_m \wedge \tau)|$$

where Y represents survival time, τ is a specified time, $Y \wedge \tau = \min(Y, \tau)$, and $\hat{Y}_m$ is the predicted median survival time.

Prediction error estimation with absolute error loss for censored data has received little attention in the literature. Tian et al. (2007) consider the absolute error loss for a specific class of generalized linear models, but they do not consider censored data and ignore variation due to estimation of parameter θ, as well as variable and model selection. We develop point and confidence interval estimators for prediction error with absolute error loss for survival models and examine the performance of these estimators through simulation studies in this chapter.

3.2 Estimators of Prediction Error and Confidence Intervals

The training data sets D are assumed of the form $D = \{(T_i, \delta_i, Z_i), i = 1, \ldots, n\}$, where $T_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \leq C_i)$, as defined in section 2.2. As there, it is assumed that survival time Y and censoring time C are conditionally independent, given covariates Z. We concentrate on the three point estimators of prediction error and the confidence intervals, as discussed in sections 1.2.1 and 2.4, combined with the IPCW adjustments given in section 2.2.1.

The model-based estimator $\hat{\pi}^m$, the estimator based on apparent loss $\hat{\pi}^A$ and the cross-validation estimator $\hat{\pi}^{cv}$ are easily used with this bounded absolute relative loss function. Since an adjustment term Ω for apparent loss is not provided by Efron's (2004) or others' results for the absolute error loss, we take $\hat{\pi}^A = \mathrm{AL}$ without an adjustment, and note that some underestimation of prediction error can be expected when n is not very large.

The model-based estimator $\hat{\pi}^m$ given by (1.12) is obtained via simulation. We first generate K sets of training data $D_k^* = \{(t_i^{*k}, \delta_i^{*k}, z_i), i = 1, \ldots, n\}$ and test data $Y_k^* = \{(y_i^{*k}, z_i), i = 1, \ldots, n\}$, $k = 1, \ldots, K$, from $F_{\hat{\theta}}(y|z)$. $\hat{\theta}$ is estimated under the chosen model family $F_\theta(y|z)$ with data D following variable selection. The values of Z are assumed known and fixed. Censoring times have to be generated for each D_k^*, and this is done using a model $\hat{S}^c(c|z)$, estimated from the censoring times in the given data D. $\hat{S}^c(c|z)$ is also used to give weights to the observed losses when $\hat{\pi}^A$ and $\hat{\pi}^{cv}$ are used with the IPCW approach. The model-based prediction error estimator is then

$$\hat{\pi}^m = \frac{1}{K}\sum_{k=1}^{K} PL_k = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n} L^\tau(y_i^{*k}, \hat{G}(z_i;\hat{\theta}_k^*)),$$

where $\hat{G}(z_i;\hat{\theta}_k^*)$ is the median of Y given $Z = z_i$, based on the model $F_{\hat{\theta}_k^*}(y|z)$ obtained from D_k^*.

The IPCW version of apparent loss (AL) is, as in (2.14)

$$\widehat{\text{AL}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{\hat{S}^c(y_i \wedge \tau | z_i)} L^\tau(y_i, \hat{y}_i). \tag{3.1}$$

As we discussed earlier, a penalty adjustment term is not available so $\hat{\pi}^A = \widehat{\text{AL}}$.

The V-fold cross-validation estimator $\hat{\pi}^{cv}$ is obtained by replacing $\hat{y}_i$ in (3.1) with $\hat{y}_{i(-v)}$, as in (2.15)

$$\hat{\pi}^{cv} = \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v} \frac{\Delta_i}{\hat{S}^c(y_i \wedge \tau | z_i)} L^\tau(y_i, \hat{y}_{i(-v)}). \tag{3.2}$$

For each of the three estimators, the confidence intervals for π are obtained using either the parametric or nonparametric bootstrap approach as described in section 2.4. Simulation results for point and confidence interval estimation are given in the next section.

3.3 Simulation Studies

A number of authors have conducted simulation studies for estimation of prediction error for survival models, mainly for the binary status variable W_t (e.g. Rosthøj and Keiding 2004; Gerds and Schumacher; 2006 & 2007; Uno et al. 2007). No one has considered confidence interval estimation except for Uno et al. (2007), and most studies ignore model selection and misspecification.

Variable	logbun	hgb	scalc	age	logpbm	logwbc	frac	protein	gender
β	-1.85	0.12	-0.16	0.02	-0.4	-0.01	-0.01	-0.01	0.01
sd(variable)	0.31	2.56	1.82	10.3	0.36	0.24		6.01	

Table 3.1: Covariates and true regression coefficients of the simulation model (3.3). logbun: log(Blood Urea Nitrogen); hgb: Hemoglobin; scalc: Serum Calcium; age: Age in years; logpbm: log(Percentage of Plasma Cells in Bone Marrow); logwbc: log(White Blood Cell Count); frac: Fractures present at diagnosis 0-no, 1-yes; protein: Proteinuria; gender: 0-male, 1-female. The empirical standard deviation of the continuous variables are also given.

We investigate point estimators and confidence intervals for the bounded absolute log relative error loss function ((2.2), page 22) with simulation studies. The effects of model misspecification and variable selection are considered. The simulation settings are based on survival data for multiple myeloma patients, which has features typical of many situations. The data set contains 65 survival times, 17 of which are censored, and 16 covariates (Krall et al. 1975). We simulated data from a 9-variable model:

$$\log Y_i = \beta_0 + \beta_1 z_{1i} + \ldots + \beta_9 z_{9i} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.3}$$

where ε_i follows a standard extreme value distribution EV(0, 1) and Y_i is measured in months. Thus the conditional distribution $F_T(y|z)$ follows an exponential distribution with mean $E(Y|Z = z_i) = \exp(\beta_0 + \beta_1 z_{1i} + \ldots + \beta_9 z_{9i})$ months. The names of the covariates Z and the true regression coefficients are given in Table 3.1. Of the nine variables, we let three (log(Blood Urea Nitrogen), Hemoglobin and Serum Calcium) have large effects, two (log(Percentage of Plasma Cells in Bone Marrow) and age) have moderate effects, and the other four variables have very small effects. This simulation model is consistent with variables found important in analysis of the data (Krall et al. 1975; Lawless and Singhal 1978).

The simulated survival times were subjected to random censoring, with censoring times generated according to the empirical censoring distribution of the multiple myeloma data set. To examine the effects of sample size, we conducted simulations with $n = 65$ and 400. Since the correlations between the nine variables in the original data set are not high, for the $n = 400$ case we generate values for each of the nine variables independently, roughly following the empirical distribution of the corresponding Z in the original data set. For example, the binary variable "Fractures" is generated from a Bernoulli distribution (Pr(having fractures) $= 0.75$), and the continuous variable "log(Blood Urea Nitrogen)" is generated from a univariate normal distribution with sample mean and standard deviation equal to those in the original data set. Once generated, these 400 z_i values were kept fixed and Y_i values were generated from model (3.3).

In order to make the simulation study feasible, we consider a simple variable selection procedure where the full model is fitted and then variables with p-values greater than or equal to 0.2 are removed, and the data refitted to the remaining variable(s), giving the final model. Two families of models, Weibull and Lognormal, are fitted to the simulated data. The Weibull corresponds to using (3.3) with ε_i having an extreme value distribution with location parameter 0 and unknown scale parameter b (see Lawless 2003, page 20), and the Lognormal model corresponds to using (3.3) with ε_i having a normal distribution with mean 0 and unknown standard deviation b. In this setting, the Weibull is the "correct" model family and the Lognormal model is an example of a misspecified model.

3.4 Simulation Results

Simulation results for the point estimators of $\pi(M; F_T)$ (1.4) are shown in Table 3.2 for the loss function (2.2) truncated at $\tau = 50$ months, at which time $\hat{S}(\tau) = 0.21$ and $\hat{S}^c(\tau) = 0.65$. To keep the computing demands reasonable while still allowing a decent comparison of methods, we used 500 simulation runs. The model-based estimators $\hat{\pi}^m$ as in (1.12) are estimated with simulation from the fitted model $F_{\hat{\theta}}$ following variable selection; here $\hat{\theta} = (\hat{\beta}, \hat{b})$ includes the regression coefficients in (3.3) and the scale parameter in the error distribution. In the present simulation setting, the estimates $\hat{\pi}^m$ were obtained from $K = 50$ simulated samples. We use $B = 100$ samples to obtain $\hat{\text{var}}(\hat{\pi}^m)$ and confidence intervals (see section 2.4) in each case. Larger values of B and K are desirable and can be used for single training samples. However, the process was repeated 500 times in the simulation study and we chose to use these smaller values while still obtaining a quite good picture of the method's properties.

For the estimator $\hat{\pi}^A$ (3.1) there was no optimism adjustment, and 5-fold cross-validation is used for $\hat{\pi}^{cv}$ (3.2). We used $B = 500$ bootstrap samples (both parametric, based on $F_{\hat{\theta}}$, and nonparametric) to obtain variance estimates. Confidence intervals for all three methods were based on treating either $\hat{\pi}$ or $\log(\hat{\pi})$ as normally distributed random variables; we also considered bias correction as well as the percentile method and the basic percentile method (see section 2.4). The averages of the prediction error estimates and their standard deviations are given in Table 3.2, and the coverage proportions of the confidence intervals are given in Tables 3.3 and 3.4, for sample size 65 and 400, respectively. Simulation data for each estimator and bootstrap method combination was generated independently. Therefore, a total of 24 independent simulations, one for each row of Table 3.2, were conducted.

We use simulations to obtain the true prediction errors (1.4) for the Weibull and Lognormal prediction procedures were obtained from simulation. We generated 10,000 pairs of training and test data sets under the true Weibull model (3.3). The training data was subjected to random censoring described in section 3.3. Either a Weibull or a Lognormal model was fitted to the training data and subjected to the variable selection scheme discussed in section 3.3,

Model fitted	Estimator	Bootstrap	Ave($\hat{\pi}$)	Esd($\hat{\pi}$)	Ave($\hat{sd}(\hat{\pi})$)
	Model based	Parametric	0.829	0.115	0.102
	$\hat{\pi}^m$	Nonparametric	0.818	0.116	0.113
Weibull	Apparent Loss,	Parametric	0.780	0.113	0.102
(0.915)	$\hat{\pi}^A$	Nonparametric	0.780	0.113	0.113
$n = 65$	5 fold CV,	Parametric	0.946	0.137	0.128
	$\hat{\pi}^{cv}$	Nonparametric	0.945	0.131	0.143
	Model based	Parametric	0.878	0.125	0.085
	$\hat{\pi}^m$	Nonparametric	0.877	0.116	0.109
Lognormal	Apparent Loss	Parametric	0.810	0.119	0.092
(0.942)	$\hat{\pi}^A$	Nonparametric	0.812	0.125	0.119
$n = 65$	5 fold CV,	Parametric	1.001	0.146	0.122
	$\hat{\pi}^{cv}$	Nonparametric	0.971	0.146	0.147
	Model based	Parametric	0.856	0.040	0.040
	$\hat{\pi}^m$	Nonparametric	0.858	0.040	0.041
Weibull	Apparent Loss	Parametric	0.853	0.046	0.044
(0.869)	$\hat{\pi}^A$	Nonparametric	0.851	0.043	0.045
$n = 400$	5 fold CV	Parametric	0.872	0.045	0.045
	$\hat{\pi}^{cv}$	Nonparametric	0.873	0.046	0.046
	Model based	Parametric	0.898	0.047	0.033
	$\hat{\pi}^m$	Nonparametric	0.900	0.050	0.047
Lognormal	Apparent Loss	Parametric	0.864	0.044	0.037
(0.883)	$\hat{\pi}^A$	Nonparametric	0.861	0.045	0.048
$n = 400$	5 fold CV	Parametric	0.887	0.047	0.039
	$\hat{\pi}^{cv}$	Nonparametric	0.888	0.047	0.049

Table 3.2: Simulation results for estimates of prediction error based on absolute error loss (500 simulations). True prediction error estimated from 10,000 simulations is given in the left column in parenthesis for each model. Ave($\hat{\pi}$) is the average and Esd($\hat{\pi}$) is the standard deviation of $\hat{\pi}$ over the 500 simulation runs; Ave($\hat{sd}(\hat{\pi})$) is the average of the standard deviation estimates for $\hat{\pi}$ over the 500 simulation runs. “Bootstrap” refers to the way the variance estimates for $\hat{\pi}$ are obtained, as described in section 2.4. $\hat{\pi}^m$, $\hat{\pi}^A$ and $\hat{\pi}^{cv}$ are given by (1.12), (3.1) and (3.2).

giving predicted median survival time $G(z_i; \hat{\theta})$. Prediction error was assessed with the test data. The true prediction errors for the two modelling procedures are given in Table 3.2. We note that the true prediction error is slightly smaller under the true Weibull family, but the Lognormal family also performs quite well. This is because the estimates of the median of Y given Z that Lognormal model produces are not too different from those of the Weibull model.

Table 3.2 indicates that, as expected, $\hat{\pi}^A$ underestimates π substantially when n is small ($n = 65$) but only a little when $n = 400$. Conversely, $\hat{\pi}^{cv}$ overestimates π somewhat when $n = 65$ but is very accurate for $n = 400$. The model-based estimator $\hat{\pi}^m$ is somewhat biased when $n = 65$, no matter which family of models is used. However, a more serious problem is that when the incorrect model (Lognormal) is used, the variance estimated with the parametric bootstrap procedure is biased downward across all methods, with the problem being most severe for the model-based estimator $\hat{\pi}^m$. This is seen by comparing the last column $\text{Ave}(\hat{sd}(\hat{\pi}))$ of Table 3.2 with the empirical standard deviation (Esd) column. For example, the empirical standard deviation of $\hat{\pi}^m$ is 0.125 when the sample size is 65 under the Lognormal model, and the average of its parametric bootstrap estimates is only 0.085. The underestimation persists in the large sample case when $n = 400$; the $\text{Esd}(\hat{\pi}^m)$ is 0.047 and the average of its estimates is 0.033. This underestimation produces low confidence interval coverages of π for the misspecified model, as seen in Table 3.3 and 3.4.

The coverage proportions for confidence intervals at three nominal levels, 0.90, 0.95 and 0.99, are summarized in Tables 3.3 and 3.4, for the two sample sizes $n = 65$ and 400, respectively. A method performs well when the coverage proportions are close to the nominal levels. With 500 simulations, the empirical coverage probability would be expected to fall in the interval 0.93 and 0.97 for 95% confidence intervals, for example. Table 3.3 shows that when the sample size is small, log-transformation of $\hat{\pi}$ improves the coverage of confidence intervals; it is seen by comparing the "for $\log(\hat{\pi})$" column with the "for $\hat{\pi}$" column. Similarly we find that the coverage is better when the bias correction is applied to $\hat{\pi}^m$ and $\hat{\pi}^A$. Overall, table 3.3 suggests

	Nominal	Normal Approximation		Bias corrected NA		Basic	Percentile
	Level	for $\hat{\pi}$	for $\log(\hat{\pi})$	for $\hat{\pi}$	for $\log(\hat{\pi})$	method	method
Weibull	0.90	0.706	0.798	0.818	0.916	0.798	0.500
(Parametric	0.95	0.776	0.880	0.886	0.952	0.852	0.580
Bootstrap, $\hat{\pi}^m$)	0.99	0.882	0.962	0.952	0.986	0.914	0.714
Weibull	0.90	0.740	0.830	0.824	0.916	0.506	0.480
(Nonparametric	0.95	0.802	0.910	0.876	0.972	0.584	0.568
Bootstrap, $\hat{\pi}^m$)	0.99	0.898	0.978	0.966	0.998	0.686	0.702
Weibull	0.90	0.576	0.702	0.724	0.826	0.714	0.354
(Parametric	0.95	0.676	0.806	0.786	0.892	0.774	0.486
$\hat{\pi}^A$)	0.99	0.812	0.938	0.892	0.960	0.856	0.686
Weibull	0.90	0.630	0.770	0.798	0.918	0.782	0.384
(Nonparametric	0.95	0.726	0.862	0.858	0.966	0.838	0.500
$\hat{\pi}^A$)	0.99	0.850	0.964	0.938	0.990	0.908	0.696
Weibull	0.90	0.864	0.916	0.678	0.694	0.686	0.874
(Parametric	0.95	0.928	0.954	0.774	0.790	0.788	0.928
$\hat{\pi}^{cv}$)	0.99	0.986	0.994	0.878	0.896	0.882	0.980
Weibull	0.90	0.932	0.954	0.750	0.780	0.762	0.914
(Nonparametric	0.95	0.966	0.976	0.838	0.858	0.842	0.958
$\hat{\pi}^{cv}$)	0.99	0.992	0.998	0.936	0.952	0.934	0.984
Lognormal	0.90	0.654	0.724	0.710	0.788	0.674	0.684
(Parametric	0.95	0.724	0.796	0.796	0.868	0.746	0.764
Bootstrap)	0.99	0.842	0.912	0.898	0.954	0.840	0.860
Lognormal	0.90	0.764	0.850	0.824	0.886	0.536	0.538
(Nonparametric	0.95	0.816	0.904	0.886	0.944	0.602	0.614
Bootstrap)	0.99	0.904	0.968	0.970	0.988	0.692	0.706
Lognormal	0.90	0.540	0.620	0.620	0.734	0.618	0.380
(Parametric	0.95	0.608	0.694	0.710	0.828	0.684	0.446
$\hat{\pi}^A$)	0.99	0.746	0.856	0.828	0.906	0.802	0.616
Lognormal	0.90	0.648	0.776	0.780	0.884	0.776	0.438
(Nonparametric	0.95	0.714	0.866	0.856	0.936	0.834	0.546
$\hat{\pi}^A$)	0.99	0.858	0.946	0.916	0.986	0.906	0.678
Lognormal	0.90	0.810	0.824	0.610	0.626	0.618	0.870
(Parametric	0.95	0.894	0.902	0.710	0.726	0.722	0.922
$\hat{\pi}^{cv}$)	0.99	0.974	0.982	0.830	0.842	0.836	0.976
Lognormal	0.90	0.884	0.940	0.724	0.748	0.722	0.886
(Nonparametric	0.95	0.948	0.970	0.810	0.840	0.818	0.938
$\hat{\pi}^{cv}$)	0.99	0.990	0.992	0.918	0.936	0.920	0.980

Table 3.3: Coverage proportions for three nominal confidence levels, 0.90, 0.95 and 0.99, based on 500 simulations for sample size $n = 65$.

that

1. $\hat{\pi}^m$: Under the Weibull model, both parametric and nonparametric procedures work well with bias correction and treating $\log(\hat{\pi}^m)$ as approximately normal. Under the Lognormal model, the nonparametric procedure still produces confidence intervals with approximately the right coverage when bias correction and $\log(\hat{\pi}^m)$ are used; but when the parametric procedure is used, the coverage is substantially lower. This is largely due to the underestimation of $\text{var}(\hat{\pi}^m)$ when parametric bootstrap is used under the incorrect model, as seen in Table 3.2. This point will be discussed further in the next section.

2. $\hat{\pi}^A$: The nonparametric bootstrap procedure with bias correction and the use of $\log(\hat{\pi}^A)$ works well for both Weibull and Lognormal models. The parametric bootstrap does not work so well, and the undercoverage is more severe under the incorrect model.

3. $\hat{\pi}^{cv}$: Under the Weibull model, the parametric bootstrap works well with the unadjusted estimate and the use of $\log(\hat{\pi}^{cv})$, the nonparametric bootstrap gives some overcoverage using the unadjusted estimate (for $\hat{\pi}^{cv}$ and $\log(\hat{\pi}^{cv})$). Under the Lognormal model, the coverage is approximately right with the nonparametric bootstrap procedure and the use of unadjusted $\hat{\pi}^{cv}$. In addition, the results suggest that bias correction does not work with $\hat{\pi}^{cv}$ under either model.

4. The basic method works poorly across the table.

5. The percentile method works well only when confidence intervals are based on $\hat{\pi}^{cv}$ and the nonparametric bootstrap.

When the sample size increases to 400, the coverage proportions are similar for $\hat{\pi}$ and $\log(\hat{\pi})$ (Table 3.4). The table suggests that

1. $\hat{\pi}^m$: Under the correct model family, both parametric and nonparametric bootstrap work well, for the normal approximation considered here. Under the incorrect model family,

	Nominal	Normal Approximation		Bias corrected NA		Basic	Percentile
	Level	for $\hat{\pi}$	for $\log(\hat{\pi})$	for $\hat{\pi}$	for $\log(\hat{\pi})$	method	method
Weibull	0.90	0.876	0.888	0.888	0.888	0.872	0.798
(Parametric	0.95	0.924	0.938	0.932	0.936	0.912	0.868
Bootstrap, $\hat{\pi}^m$)	0.99	0.968	0.984	0.978	0.990	0.966	0.926
Weibull	0.90	0.912	0.920	0.898	0.902	0.896	0.876
(Nonparametric	0.95	0.954	0.956	0.946	0.954	0. 950	0.922
Bootstrap, $\hat{\pi}^m$)	0.99	0.984	0.988	0.982	0.984	0.974	0.974
Weibull	0.90	0.860	0.870	0.786	0.794	0.782	0.834
(Parametric	0.95	0.900	0.916	0.856	0.868	0.852	0.884
$\hat{\pi}^A$)	0.99	0.964	0.972	0.938	0.946	0.920	0.948
Weibull	0.90	0.868	0.874	0.884	0.908	0.882	0.838
(Nonparametric	0.95	0.914	0.934	0.942	0.956	0.938	0.892
$\hat{\pi}^A$)	0.99	0.982	0.986	0.984	0.988	0.984	0.966
Weibull	0.90	0.884	0.890	0.778	0.780	0.774	0.930
(Parametric	0.95	0.946	0.950	0.846	0.854	0.846	0.960
$\hat{\pi}^{cv}$)	0.99	0.996	1.000	0.944	0.946	0.940	0.992
Weibull	0.90	0.904	0.906	0.880	0.880	0.880	0.894
(Nonparametric	0.95	0.958	0.956	0.926	0.928	0.926	0.956
$\hat{\pi}^{cv}$)	0.99	0.990	0.992	0.988	0.986	0.982	0.988
Lognormal	0.90	0.736	0.746	0.684	0.696	0.654	0.734
(Parametric	0.95	0.834	0.836	0.776	0.766	0.752	0.816
Boostrap)	0.99	0.924	0.922	0.890	0.890	0.852	0.890
Lognormal	0.90	0.872	0.868	0.822	0.810	0.800	0.800
(Nonparametric	0.95	0.942	0.940	0.906	0.900	0.886	0.878
Boostrap)	0.99	0.982	0.980	0.970	0.968	0.942	0.938
Lognormal	0.90	0.800	0.802	0.658	0.660	0.648	0.812
(Parametric	0.95	0.882	0.880	0.750	0.758	0.740	0.890
$\hat{\pi}^A$)	0.99	0.946	0.952	0.880	0.886	0.856	0.954
Lognormal	0.90	0.858	0.870	0.888	0.900	0.878	0.826
(Nonparametric	0.95	0.910	0.928	0.930	0.948	0.920	0.874
$\hat{\pi}^A$)	0.99	0.974	0.988	0.988	0.994	0.980	0.954
Lognormal	0.90	0.824	0.822	0.712	0.704	0.712	0.806
(Parametric	0.95	0.892	0.890	0.804	0.800	0.800	0.872
$\hat{\pi}^{cv}$)	0.99	0.976	0.968	0.922	0.918	0.900	0.960
Lognormal	0.90	0.926	0.924	0.886	0.880	0.886	0.918
(Nonparametric	0.95	0.960	0.964	0.944	0.944	0.952	0.958
$\hat{\pi}^{cv}$)	0.99	0.988	0.990	0.986	0.990	0.982	0.986

Table 3.4: Coverage proportions for three nominal confidence levels, 0.90, 0.95 and 0.99, based on 500 simulations for sample size $n = 400$.

however, the coverage proportions are close to nominal levels only when the nonparametric bootstrap is used with the unadjusted estimator.

2. $\hat{\pi}^A$: Under both models, the nonparametric bootstrap procedure with bias corrected estimator works well.
3. $\hat{\pi}^{cv}$: Under the Weibull model, both parametric and nonparametric work well with the unadjusted estimators, log-transformed or not. Under the Lognormal model, the only method that gives satisfactory coverage proportions is the nonparametric bootstrap procedure with the unadjusted estimators.
4. The basic method gives approximately right coverage when the nonparametric procedure is used under the Weibull model. Under the incorrect Lognormal model, it only works well with nonparametric bootstrap based on $\hat{\pi}^{cv}$.
5. The percentile method only works with nonparametric bootstrap based on $\hat{\pi}^{cv}$, similar to the results seen in Table 3.3.

To summarize, the parametric bootstrap procedure for the construction of confidence intervals is susceptible to model misspecification, therefore it is not recommended. The nonparametric bootstrap based on the cross-validation estimator $\hat{\pi}^{cv}$ works well for all scenarios investigated. The estimator $\hat{\pi}^A$, or $\log(\hat{\pi}^A)$ for small sample size, can also be used with the nonparametric bootstrap to correct for bias and for confidence interval estimation.

The choice of $\tau = 50$ months in the simulation study is arbitrary, we believe that the results should be similar for other choices of τ, but more simulation studies are needed to verify it.

3.5 Problems with the Model-based Approach

We note in Table 3.2 that the parametric bootstrap procedure fails to produce reliable estimators for var($\hat{\pi}$) in the face of model misspecification. Simple calculations show that the degree of undercoverage of confidence intervals seen in Tables 3.3 and 3.4 under normal approximation can be explained by the bias of the estimator and the degree of underestimation of var($\hat{\pi}$) seen in Table 3.2. We illustrate the connection between the underestimation of variance and the undercoverage of confidence interval with a normal example. Problems with the model-based procedure for the estimation of $\hat{\pi}$ and of $\hat{\text{var}}(\hat{\pi})$ under a misspecified model are discussed later in the section.

To obtain a confidence interval for π, we need to know the distribution of $\hat{\pi}$. Let $\hat{sd}(\hat{\pi})$ denote the estimate of $sd(\hat{\pi})$, the standard deviation of $\hat{\pi}$, and define the "sd factor"

$$sd.f = E[\hat{sd}(\hat{\pi})]/sd(\hat{\pi})$$

and the "bias factor",

$$b.f = (E_{F_T}(\hat{\pi}) - \pi)/sd(\hat{\pi}).$$

$sd.f$ measures the under or over estimation of standard deviation of $\hat{\pi}$ and $b.f$ measures the bias of $\hat{\pi}$ relative to the variance of $\hat{\pi}$. Assuming $\hat{\pi}$ is approximately normally distributed, the effects of the sd factor and bias factor on the coverage probability can be explained with Figure 3.1, which displays two normal density functions. Suppose the lighter curve N($\mu = 0.5, \sigma^2 = 1.2^2$) represents the true distribution of $\hat{\pi}$, so the confidence interval based on this curve would have the right coverage. But the true curve is unknown to us and suppose instead the darker curve N(0, 1) is estimated from the data and used to construct the confidence interval for π. The two dashed vertical lines at -1.645 and 1.645 give a nominal 90% confidence interval under the darker curve. The true coverage probability, however, is given by the area under the lighter curve between lines -1.645 and 1.645, which corresponds to the area under the darker curve

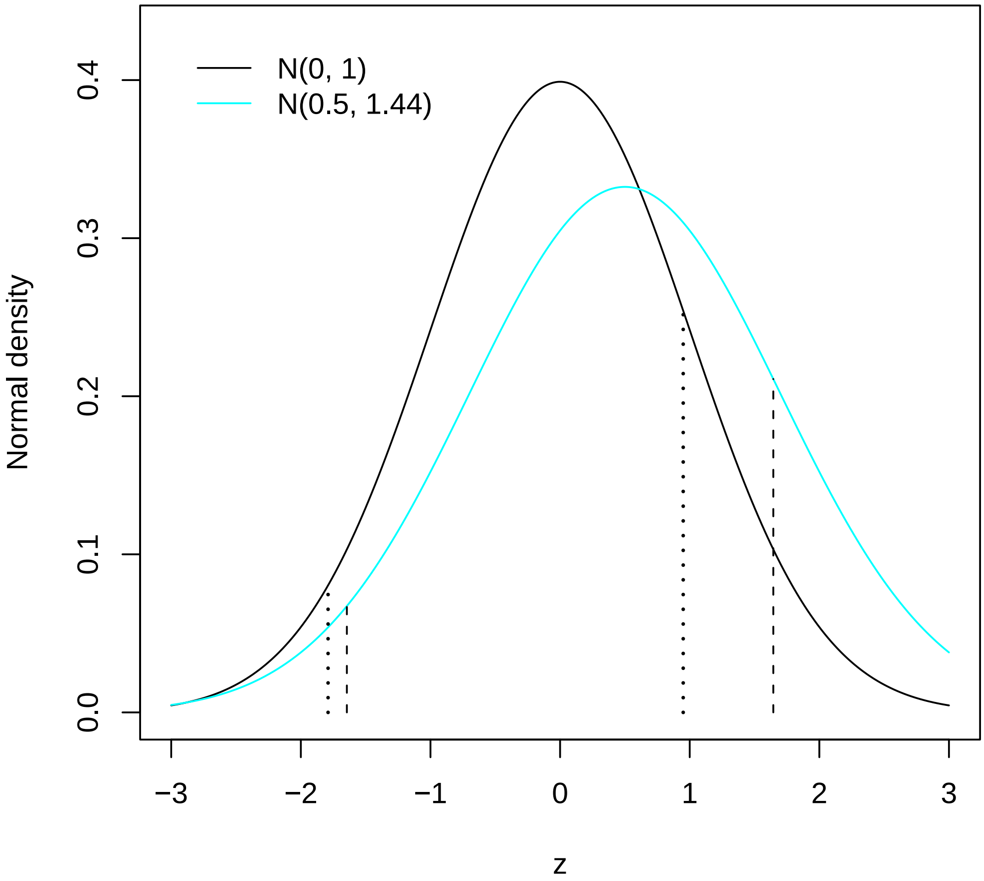

Figure 3.1: The effects of bias and underestimation of variance of $\hat{\pi}$ on the coverage probability.

between the two dotted vertical lines, of which the x-coordinates are

$$-1.645\frac{1}{1.2}+\frac{0-0.5}{1.2}=-1.79,\ 1.645\frac{1}{1.2}+\frac{0-0.5}{1.2}=0.95,$$

where $sd.f = 1/1.2 = 0.83$, and $b.f = (0 - 0.5)/1.2 = -0.42$. Hence, the true coverage probability is $\Phi(0.95) - \Phi(-1.79) = 79.3\%$. Since the sd factor operates multiplicatively and the bias factor is not very big, the reduction in coverage probability is largely due to the sd factor. The underestimation of the standard deviation by 17% alone reduces the coverage probability from 90% to 82.9%.

In the simulation study above, we find that the variance underestimation of $\hat{\pi}$ is the main reason for the undercoverage observed, especially for the misspecified Lognormal model. We calculated the coverage proportions with the data in Table 3.2 and the method described above, and found the results agree well with the observed coverage probabilities seen in Tables 3.3 and 3.4. For example, let us look at the estimate $\hat{\pi}^A$ under the Lognormal model and sample size 400 (Table 3.2). The true prediction error is $\pi = 0.883$ and $E(\hat{\pi}) \doteq \bar{\hat{\pi}}^A = 0.864$, the standard deviation is $sd(\hat{\pi}) \doteq Esd(\hat{\pi}^A) = 0.044$ and the average estimate from the parametric bootstrap procedure gives $E(\hat{sd}(\hat{\pi})) \doteq 0.037$. Therefore, approximately,

$$sd.f = 0.037/0.044 = 0.84, \qquad b.f = (0.864 - 0.883)/0.044 = -0.43.$$

For nominal 90%, the coverage probability under normal assumption is

$$\Phi(1.645 \times 0.84 - 0.43) - \Phi(-1.645 \times 0.84 - 0.43) = 79.4\%.$$

Similar calculation shows that the coverage probabilities for nominal levels 95% and 99% are 86.9% and 95.4%, respectively. These values agree with the observed coverage probabilities 80.0%, 88.2% and 94.6% in Table 3.4, under the column “Normal Approximation for $\hat{\pi}$” and row “Lognormal (Parametric $\hat{\pi}^A$)”.

The confidence interval coverage problems with model-based estimation are largely due to biases in the estimation of $\text{var}(\hat{\pi})$ or, more generally, in the distribution of $\hat{\pi}$, when an incorrect model is used to generate D_b^* for the estimation of var($\hat{\pi}$). A misspecified model can return similar point predictors $\hat{Y}$ to a correctly specified model when predictors are measures of location such as medians or means. This is supported by our simulation results: the true prediction errors π are not that different under the correct and incorrect models (0.915 and 0.942 for $n = 65$, and 0.869 and 0.883 for $n = 400$, see Table 3.2). However, the models can differ substantially in the distribution tails, leading to bias in estimates of measures of variation and to estimates of var($\hat{\pi}$) when a model-based procedure is used, such as $\hat{\pi}^m$ and var($\hat{\pi}$) estimated by the parametric bootstrap.

In the present simulation setting, $\hat{\pi}^m$ is not very biased under the incorrect model, but $\hat{\text{var}}(\hat{\pi})$ is seriously biased when the parametric bootstrap procedure is used under the incorrect model. Nonetheless, we can easily find an example where the model-based estimator $\hat{\pi}^m$ differs substantially in mean from π under model misspecification: when we reverse the role of Lognormal and Weibull models in the present simulation setting by generating data from the "true" Lognormal model and fit the data with the "misspecified" Weibull model. In particular, suppose we simulate the data from the model

$$\log Y_i = \beta_0 - \gamma + \beta_1 z_{1i} + \ldots + \beta_9 z_{9i} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.4}$$

where $\gamma = 0.5772...$ is the Euler constant, ε_i follows a normal distribution N(0, $\pi^2/6$) and Y_i is measured in months. Here $\pi = 3.14159...$ refers to the mathematical constant, sometimes known as the circular constant. The regression coefficients β for covariates are kept unchanged; their values are given by Table 3.1. γ is added to the intercept β_0 so that the mean $E(\log(Y)|Z)$ under the model (3.4) is the same as $E(\log(Y)|Z)$ under the model (3.3). The variance $\sigma^2 = \pi^2/6$ is used so that the variance of the conditional distribution $Y|Z$ remains unchanged. Both Weibull and Lognormal models are fit to the simulated data with the same variable selection procedure described in section 3.3. Note that in this case Lognormal model is the true distribution and the Weibull is an example of misspecified model.

Simulation results for the point estimators of prediction error are shown in Table 3.5 and

Model fitted	Estimator	Ave($\hat{\pi}$)	Esd($\hat{\pi}$)	Ave($\hat{sd}(\hat{\pi})$)	
				PBS	NPBS
Weibull	$\hat{\pi}^m$	0.996	0.047	0.047	0.044
$n = 400$	$\hat{\pi}^A$	0.868	0.034	0.051	0.037
(0.894)	$\hat{\pi}^{cv}$	0.897	0.037	0.053	0.039
Lognormal	$\hat{\pi}^m$	0.869	0.032	0.031	0.032
$n = 400$	$\hat{\pi}^A$	0.855	0.035	0.036	0.037
(0.880)	$\hat{\pi}^{cv}$	0.882	0.037	0.038	0.039

Table 3.5: Simulation results for estimates of prediction error based on absolute error loss for model (3.4) (500 simulations). True prediction error estimated from 10,000 simulations is given in the left column in parenthesis for each model. Ave($\hat{\pi}$) is the average and Esd($\hat{\pi}$) is the standard deviation of $\hat{\pi}$ over the 500 simulation runs; Ave($\hat{sd}(\hat{\pi})$) is the average of the standard deviation estimates for $\hat{\pi}$ over the 500 simulation runs. PBS and NPBS refer to parametric and nonparametric bootstrap estimation of var($\hat{\pi}$), respectively, as described in section 2.4.

for the confidence interval procedures in Table 3.6, for the same loss function (2.2) truncated at $\tau = 50$ months and sample size $n = 400$.

Table 3.5 shows that

1. The model-based estimator $\hat{\pi}^m$ has a large bias when the incorrect Weibull model is used (average $(\hat{\pi}^m) = 0.996$, when the true value is 0.894). This results in very poor confidence interval coverage even for the nonparametric bootstrap procedure (0.404 for nominal 0.95, see Table 3.6).

2. The model-based parametric bootstrap procedure overestimates var($\hat{\pi}^A$) and var($\hat{\pi}^{cv}$) substantially under the incorrect Weibull model, producing the overcoverage of confidence intervals seen in Table 3.6.

3. The nonparametric bootstrap procedure provides satisfactory confidence interval when

Model fitted	Method	Normal Approximation		Bias corrected NA	
		for $\hat{\pi}$	for $\log(\hat{\pi})$	for $\hat{\pi}$	for $\log(\hat{\pi})$
Weibull $n = 400$	$\hat{\pi}^m$ PBS	0.462	0.440	0.338	0.314
	$\hat{\pi}^m$ NPBS	0.404	0.388	0.200	0.196
	$\hat{\pi}^A$ PBS	0.976	0.962	0.318	0.294
	$\hat{\pi}^A$ NPBS	0.894	0.914	0.932	0.954
	$\hat{\pi}^{cv}$ PBS	0.996	0.994		
	$\hat{\pi}^{cv}$ NPBS	0.962	0.964		
Lognormal $n = 400$	$\hat{\pi}^m$ PBS	0.922	0.934	0.942	0.942
	$\hat{\pi}^m$ NPBS	0.928	0.934	0.944	0.948
	$\hat{\pi}^A$ PBS	0.884	0.896	0.850	0.840
	$\hat{\pi}^A$ NPBS	0.892	0.912	0.930	0.952
	$\hat{\pi}^{cv}$ PBS	0.946	0.950		
	$\hat{\pi}^{cv}$ NPBS	0.958	0.960		

Table 3.6: Coverage proportions for nominal 0.95 confidence levels with different methods based on 500 simulations for model (3.4). PBS and NPBS refer to the parametric and nonparametric bootstrap procedures used in the estimation of $\text{var}(\hat{\pi})$, respectively.

used with $\hat{\pi}^A$ and $\hat{\pi}^{cv}$ even under the incorrect Weibull model.

Table 3.6 confirms that the model-based estimator $\hat{\pi}^m$ can be serious biased. In section 2.3 we noted that under model misspecification $\hat{\pi}^m$ converges to $\pi(M; F_{\theta^*})$, rather than to the true prediction error $\pi(M; F_T)$. When the Lognormal model is fitted to data generated by an Exponential (Weibull) model (3.3), the corresponding $\pi(M; F_{\theta^*})$ is close to the true error of interest $\pi(M; F_T)$. This is because $\pi(M; F_T)$ is a variation measure and an incorrectly assumed normal distribution still produces a consistent estimate of variance. But this is not true in general as Table 3.5 indicates.

Both simulation studies show that the estimator for var$(\hat{\pi})$ based on the parametric bootstrap procedure is problematic under the misspecified model. For models in the location-scale family, the variance of the distribution is determined by the scale parameter b. Thus, var$(\hat{\pi})$ is closely related to the variance of the estimated scale parameter, var$(\hat{b})$. Under a misspecified model, the regression coefficients $\hat{\beta}$ (except for the intercept term $\hat{\beta}_0$ and scale parameter b) are consistently estimated when there is no censoring, but the variances of $\hat{\beta}$, including that of the scale parameter, are not correctly estimated by the maximum likelihood estimation method (Gould and Lawless 1988). For example, when data is generated from the model (3.3) and fitted with a Lognormal model with no variable selection, the average $\hat{sd}(\hat{b})$ (estimated from the information matrix for sample size 400) is 0.040 over 500 simulations, which is smaller than Esd$(\hat{b})$ = 0.065, the empirical standard deviation of $\hat{b}$. Therefore, with the parametric bootstrap procedure, var$(\hat{\sigma}^2)$ is underestimated under the incorrect Lognormal model.

A small simulation experiment confirmed the above observation. In this experiment, data sets were simulated from model (3.3) with sample size 400 and no censoring. The modelling procedure included variable selection and the loss function $L(Y, \hat{Y}) = (\log Y - \log \hat{Y})^2$ was considered. The prediction error π is then a function of σ^2, the variance of the error distribution. Table 3.7 reports the results from 500 simulations. We find the true prediction error and $E(\hat{\pi}^m)$ under the incorrect Lognormal model are very close, as expected. But, var$(\hat{\pi}^m)$ is severely

	True	Ave($\hat{\pi}^m$)	Esd($\hat{\pi}^m$)	$\hat{sd}(\hat{\pi}^m)$
Weibull	1.670	1.637	0.134	0.128
Lognormal	1.689	1.683	0.174	0.120

Table 3.7: Simulation results for estimates of prediction error based on squared error loss (500 simulations). True prediction error estimated from 10,000 simulations. Ave($\hat{\pi}^m$) is the average and Esd($\hat{\pi}^m$) is the standard deviation of $\hat{\pi}^m$ over the 500 simulation runs; Ave($\hat{sd}(\hat{\pi}^m)$) is the average of the standard deviation estimates for $\hat{\pi}^m$ using parametric bootstrap over the 500 simulation runs.

underestimated: average $\hat{sd}(\hat{\pi}^m) = 0.12$, when empirical standard deviation is 0.174.

3.6 Analytic Results under Correct and Misspecified Models

In section 3.3, we generate data from simulation model (3.3) with the error following a EV(0, 1) distribution, denoted by $f_{\theta_0}(y|z)$. When we fit a Lognormal model to the simulated data, the working model is of the form (3.3) with the error having a normal distribution, denoted by $g_\theta(y|z)$. Since $g_\theta(y|z)$ is a regular model and satisfies the regularity conditions stated in White (1982), $\hat{\theta} \to \theta^*$ in probability as sample size approaches infinity, as discussed in section 1.1.3. In the case above, the unique least false parameter θ^* can be obtained analytically by minimizing the Kullback-Leibler divergence $D(\theta)$ given in (1.8).

For the simplicity of notation, we suppress z in the following development. By definition,

$$D(\theta) = E\left\{\log\left[\frac{f_{\theta_0}(y)}{g_\theta(y)}\right]\right\} = \int f_{\theta_0}(y)\log\left[\frac{f_{\theta_0}(y)}{g_\theta(y)}\right]dy,$$

where $f_{\theta_0}(y)$ is the density function of an extreme value distribution with parameter $\theta_0 = (u, b) = (0, 1)$, and $g_\theta(y)$ is a normal density function with parameter $\theta = (\mu, \sigma^2)$. We now show that $D(\theta)$ is minimized by $(\mu^*, \sigma^{*2}) = (-\gamma, \pi^2/6)$. Note that in this section, $\gamma = 0.5772...$ refers to the Euler constant and $\pi = 3.14159...$ refers to the circular constant

Consider

$$\begin{aligned}
D(\mu,\ \sigma) &= \int \log\left[\frac{f_{\theta_0}(y)}{g_\theta(y)}\right] f_{\theta_0}(y)dy \\
&= \int \log \frac{\exp(y-\exp(y))}{\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y-\mu)^2}{2\sigma^2})} \exp(y-\exp(y))dy \\
&= \int [(\log(\sqrt{2\pi}\sigma) + (y-\exp(y) + \frac{(y-\mu)^2}{2\sigma^2})] \exp(y-\exp(y))dy.
\end{aligned}$$

Let $x = \exp(y)$,

$$\begin{aligned}
D(\mu,\ \sigma) &= \log(\sqrt{2\pi}\sigma) + (1-\frac{\mu}{\sigma^2})\int \log x \exp(-x)dx - \int x\exp(-x)dx \\
&\quad + \frac{1}{2\sigma^2}\int (\log x)^2 \exp(-x)dx + \frac{\mu^2}{2\sigma^2}\int \exp(-x)dx \\
&= \log(\sqrt{2\pi}\sigma) + (1-\frac{\mu}{\sigma^2})a - 1 + \frac{b}{2\sigma^2} + \frac{\mu^2}{2\sigma^2},
\end{aligned}$$

where $a = \int \log x \exp(-x)dx = -\gamma$, and $b = \int (\log x)^2 \exp(-x)dy = 1.978....$ To minimize $D(\mu,\ \sigma)$ with respect to μ and σ, we take partial derivatives, set them to 0 and solve for $(\mu^*,\ \sigma^*)$ with constraint $\sigma > 0$,

$$\begin{aligned}
\frac{\partial D(\mu,\ \sigma)}{\partial \mu} &= \frac{\mu - a}{\sigma^2} = 0 \\
\frac{\partial D(\mu,\ \sigma)}{\partial \sigma} &= \frac{1}{\sigma^3}(\sigma^2 - (b + \mu^2 - 2a\mu)) = 0,
\end{aligned}$$

which gives $\mu^* = a = -\gamma$ and $\sigma^{*2} = b - a^2 = \pi^2/6$.

The procedures above give a general approach for obtaining θ^*. In the case where $g_\theta(y)$ is a normal density, a straightforward solution exists for (μ^*, σ^{*2}),

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \longrightarrow \mu^* = E_{F_{\theta_0}}(Y) = -\gamma,$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \longrightarrow \sigma^{*2} = \mathrm{var}(Y) = \pi^2/6,$$

as $n \to \infty$. If $g_\theta(y)$ is not a normal density, however, there may not be a simple way to identify θ^*. For example, let $f_{\theta_0}(y)$ be a normal density with $\theta_0 = (\mu_0, \sigma_0^2) = (-\gamma, \pi^2/6)$ and $g_\theta(y)$ be an extreme value density indexed by (u, b), we want to find the parameter (u^*, b^*) that provides the best approximation to the normal distribution determined by $f_{\theta_0}(y)$. This setting corresponds to the simulation model (3.4) considered in section 3.5 with the Weibull model fitted. Then the Kullback-Leibler distance $D(u, b)$ is

$$D(u,\ b) = \int \log\left[\frac{\frac{1}{\sqrt{2\pi}\sigma_0}\exp(-\frac{(y-\mu_0)^2}{2\sigma_0^2})}{\frac{1}{b}\exp(\frac{y-u}{b} - \exp(\frac{y-u}{b}))}\right]\frac{1}{\sqrt{2\pi}\sigma_0}\exp(-\frac{(y-\mu_0)^2}{2\sigma_0^2})dy.$$

$D(u,\ b)$ cannot be evaluated analytically and is approximated by the Gauss-Hermite formula, which is then minimized with respect to u and b. The detailed steps are omitted.

Figure 3.2 shows a plot of $\frac{\partial D}{\partial u}$ and $\frac{\partial D}{\partial b}$ for different $(u,\ b)$ values. The $\bullet$ in Figure 3.2 corresponds to the least false parameter θ^*, which is given by the intersection of two lines, $\frac{\partial D}{\partial u} = 0$ and $\frac{\partial D}{\partial b} = 0$. From the graph, the u^* and b^* are approximately 0.064 and 1.28, respectively.

Taken together, N$(-\gamma,\ \pi^2/6)$ is the best approximating distribution to EV(0, 1) among all possible normal distributions. EV$(0.064,\ \pi/\sqrt{6})$ is the best approximating distribution to N$(-\gamma,\ \pi^2/6)$ among all extreme value distributions. Figure 3.3 plots the three density functions in the same graph.

Given knowledge of the three distributions, we can then obtain the theoretical true prediction errors for $n \to \infty$ under specific loss functions. Assuming Z is uniformly distributed on the 400 z_i-values generated for the $n = 400$ case (section 3.3), and that the absolute error loss of the log survival time truncated at 50 months is used, Table 3.8 gives the true prediction errors under the three error distributions.

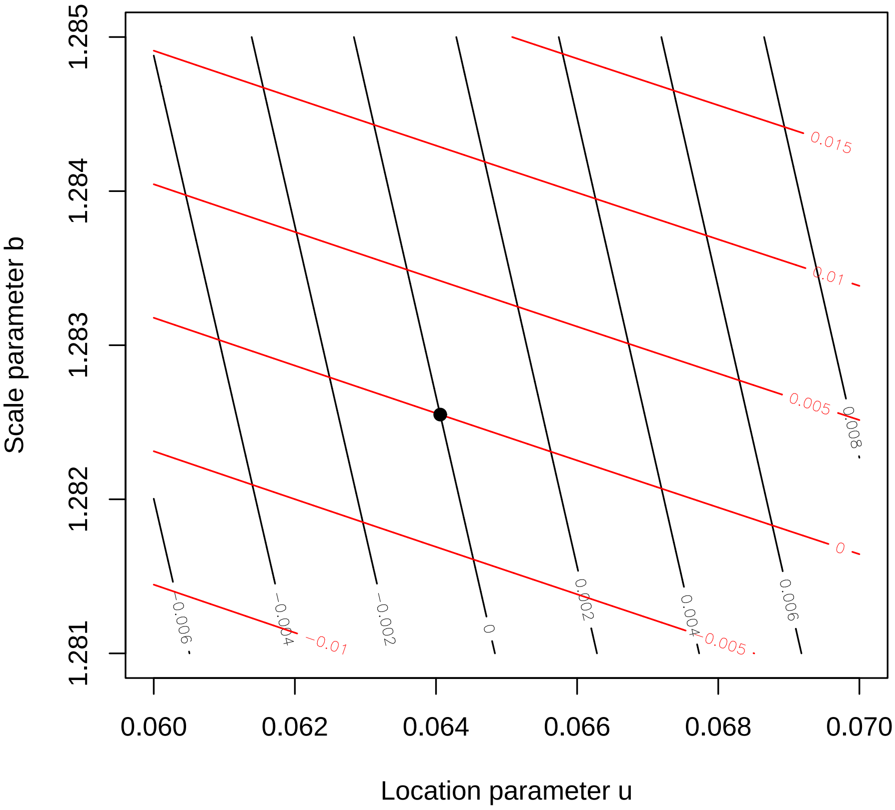

Figure 3.2: Plot of $\frac{\partial D}{\partial u}$ and $\frac{\partial D}{\partial b}$. D stands for the Kullback-Leibler distance of $\mathrm{EV}(u, b)$ to $\mathrm{N}(-\gamma, \pi^2/6)$. The black dot at the intersection of lines $\frac{\partial D}{\partial u} = \frac{\partial D}{\partial b} = 0$ gives the u^* and b^* that minimizes D.

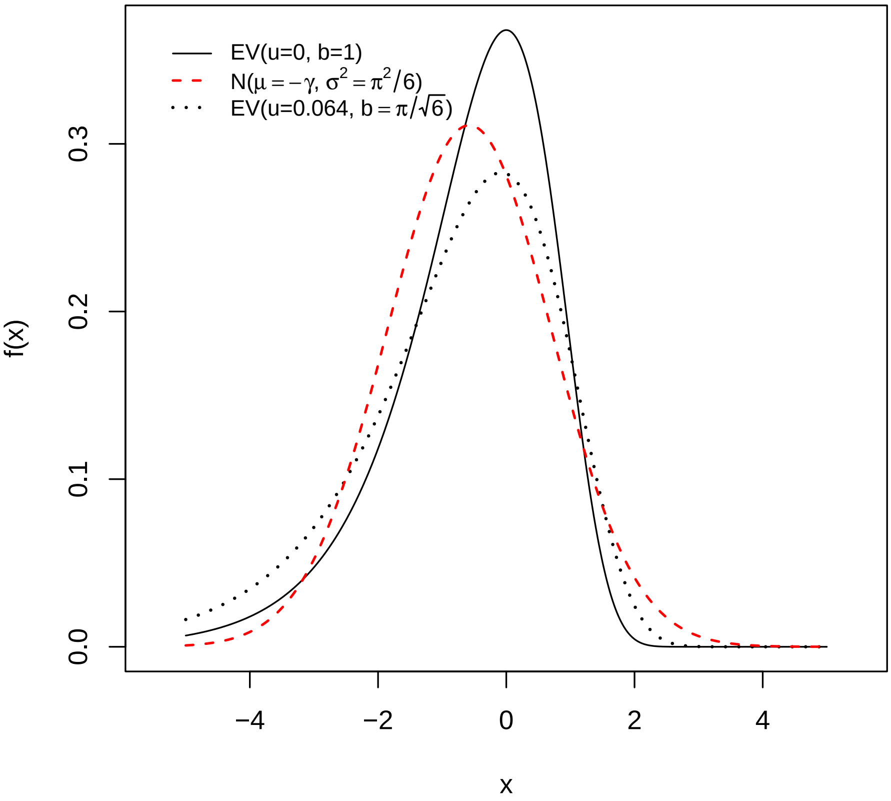

Figure 3.3: Three probability density functions. The solid, dashed and dotted curves are EV(0, 1), N($-\gamma, \pi^2/6$), and EV($0.064, \pi/\sqrt{6}$), respectively.

		True Model		
	$E_{F_T}\|L^\tau(\log(Y), \log(\hat{Y}))\|$	EV(0, 1)	N($-\gamma,\ \pi^2/6$)	EV($0.0641,\ \pi/\sqrt{6}$)
Fitted	EV(0, 1)	0.860	0.881	
	N($-\gamma,\ \pi^2/6$)	0.874	0.868	
Model	EV($0.064,\ \pi/\sqrt{6}$)		0.876	1.081

Table 3.8: Prediction error based on absolute relative error loss under correct and misspecified models for Z distributed according to the empirical Z distribution of sample size 400.

		True Model		
	$E_{F_T}\lvert L^{\tau}(\log(Y), \log(\hat{Y}))\rvert$	EV(0, 1)	N($-\gamma$, $\pi^2/6$)	EV(0.0641, $\pi/\sqrt{6}$)
Fitted	EV(0, 1)	0.847	0.863	
	N($-\gamma$, $\pi^2/6$)	0.856	0.850	
Model	EV(0.064, $\pi/\sqrt{6}$)		0.859	1.058

Table 3.9: Prediction error based on absolute error loss under correct and misspecified models for Z distributed according to the empirical Z distribution of sample size 65.

The first column in Table 3.8 corresponds to the simulation model (3.3) in section 3.3, where the error distribution of $\log(Y)$ given Z is EV(0, 1). When a Weibull model is fitted, the prediction error converges to 0.860, and the simulation result is 0.869 when $n = 400$ (Table 3.2). When a Lognormal model is fitted, the prediction error converges to 0.874, and the simulation result is 0.883 when $n = 400$ (Table 3.2). The simulation results are slightly larger than the theoretical results, which is expected. The sample size, censoring, and variable selection procedure all contribute to this difference.

The second column in Table 3.8 corresponds to the simulation model (3.4) in section 3.5, where the error distribution is N($-\gamma$, $\pi^2/6$). When a Weibull model is fitted, the best approximating error distribution is EV(0.064, $\pi^2/6$), thus the prediction error converges to 0.876, and the simulation result is 0.894 when $n = 400$ (Table 3.5). When a Lognormal model is fitted, the prediction error converges to 0.868, and the simulation result is 0.880 when $n = 400$ (Table 3.5).

A similar calculation gives the theoretical prediction errors for Z uniformly distributed on the 65 z_i-values of the multiple myeloma data, which are shown in Table 3.9. A comparison of these values to those of Table 3.2 shows that the contributions of variable selection and parameter estimation to the prediction error are greater when sample size is smaller (theoretical prediction error 0.847 vs. simulation prediction error 0.915 for Weibull model; and theoretical prediction error of 0.858 vs. simulation prediction error of 0.942 for Lognormal model).

3.7 Variable Selection

The variable selection procedure drops variables with $p > 0.2$, thus introduces variability in the final models. That is, different simulation data sets may yield different final models. There may be differences in the number of variables as well as variables themselves. Each of the nine variables in Table 3.1 has a probability of being included in the final model, depending on the model family and sample size n. The inclusion proportions for the nine variables from 500 simulation are summarized in Table 3.10. Table 3.10 shows that the inclusion proportions of the variables are different between Weibull and Lognormal models when data are generated by simulation model (3.3), especially when sample size is small. In addition, the variables with large effects are more likely to be included when sample size is large.

The confidence interval procedures, discussed in section 2.4, use either the parametric bootstrap samples simulated from the final model $F_{\hat{\theta}}(y|z)$ or the nonparametric bootstrap samples taken from D. The parametric bootstrap generates data from $F_{\hat{\theta}}(y|z)$, which varies in model size and variables included from one simulation data set to another. We summarize the results for model size in Table 3.11, which are from 500 simulations. $B = 500$ is used for parametric and nonparametric bootstrap in each simulation. Table 3.11 shows that

- Larger sample size tends to give larger model.
- The model size is slightly larger for Weibull models. The difference in model sizes between the two model families may be due to the fact that the data are generated by the extreme value distribution which has a shorter right tail than the normal distribution.
- Models based on nonparametric bootstrap samples are slightly larger than those based on parametric bootstrap samples.
- Variability in model size decreases as sample size increases, as suggested by the standard deviation of the model size.

Variable	logbun	hgb	scalc	age	logpbm	logwbc	frac	protein	gender
β	-1.85	0.12	-0.16	0.02	-0.4	-0.01	-0.01	-0.01	0.01
Weibull, $n = 65$	0.990	0.750	0.790	0.540	0.384	0.268	0.240	0.294	0.244
Weibull, $n = 400$	1.000	1.000	1.000	0.994	0.904	0.212	0.200	0.478	0.218
Lognormal, $n = 65$	0.970	0.652	0.666	0.444	0.374	0.234	0.224	0.266	0.210
Lognormal, $n = 400$	1.000	0.994	1.000	0.978	0.840	0.208	0.242	0.396	0.232

Table 3.10: The inclusion proportions for the nine covariates of simulation model (3.3) in the final model, based on 500 simulations. The true regression coefficients are also shown.

Model fitted	Sample size	Aver(model size)	PBS ($\bar{sd}$)	NPBS ($\bar{sd}$)
Weibull	$n = 65$	4.5 (1.29)	4.9 (1.23)	5.2 (1.28)
	$n = 400$	6.0 (0.94)	6.2 (0.90)	6.4 (0.96)
Lognormal	$n = 65$	4.0 (1.28)	4.5 (1.26)	4.8 (1.37)
	$n = 400$	5.9 (0.97)	6.0 (0.94)	6.2 (1.00)

Table 3.11: Simulation results for average model size when different model families are used for sample size $n = 65$ and 400, respectively. PBS refers to parametric bootstrap and NPBS to nonparametric bootstrap, which are used for confidence interval estimation. The numbers in the parenthesis give the standard deviation. $\bar{sd}$ refers to the averaged standard deviation of model size for the bootstrap procedure.

Chapter 4

Estimation of Prediction Error with Binary Loss

It is very often impossible to predict the survival time Y accurately for most individuals (e.g. Henderson et al. 2001; Henderson and Keiding 2005), but it is sometimes possible to more accurately predict who will survive beyond some time t; see for example, Korn and Simon (1990), Schemper and Henderson (2000), Rosthøj and Keiding (2004) and Henderson and Keiding (2005). In this chapter, we consider the binary survival status at a specified time t, defined as

$$W_t = I(Y > t). \tag{4.1}$$

$W_t = 0$ if subject died before time t and 1 if subject survived beyond time t. The point predictor for W_t is $\hat{W}_t$, which is also binary.

4.1 Binary Loss for Survival Status at Time t

We consider the 0-1 loss function $L(W_t, \hat{W}_t)$ for predicted survival status at time t.

$$L(W_t, \hat{W}_t) = I(W_t \neq \hat{W}_t) = |W_t - \hat{W}_t|. \tag{4.2}$$

The 0-1 loss is minimized by $\hat{W}_t = I(S_T(t|z) > 0.5)$, where $S_T(t|z) = Pr(Y > t \mid z) = 1 - F_T(t \mid z)$ is the survivor function of Y given z.

The 0-1 loss has been considered by Henderson and Keiding (2005) in a data analysis. They handled the censored observations by omitting them from the data set. Other authors considered $\hat{W}_t = S_{\hat{\theta}}(t|z)$ as a predictor for W_t with some other loss functions (e.g. Graf et al. 1999, Schemper and Henderson 2000, and Rosthøj and Keiding 2004). We restrict $\hat{W}_t$ to be 0 or 1 and use the 0-1 loss to emphasize the ability of the predictor to classify individuals correctly.

We consider the estimation of prediction error for the survival status $W_t = I(Y > t)$ with the 0-1 misclassification error loss function. By definitions (1.1) and (1.2) we have in this case,

$$\begin{aligned}\pi_{1t}(M; F_T, z, D) &= E_{W_t}[I(W_t \neq \widehat{W_t}) \mid Z = z, D] \\ &= Pr(W_t = 1 \mid z)I(\widehat{W_t} = 0) + Pr(W_t = 0 \mid z)I(\widehat{W_t} = 1).\end{aligned}$$

The equation follows because W_t is independent of $\hat{W}_t$. Then also

$$\begin{aligned}\pi_{2t}(M; F_T, z) &= E_D(\pi_{1t}) \\ &= Pr(W_t = 1 \mid z)Pr(\widehat{W_t} = 0 \mid z) + Pr(W_t = 0 \mid z)Pr(\widehat{W_t} = 1 \mid z) \\ &= S_T(t|z)(1 - Pr(S_{\hat{\theta}}(t|z) \leq 0.5)) \\ &\quad +(1 - S_T(t|z))Pr(S_{\hat{\theta}}(t|z) > 0.5),\end{aligned}$$

where $S_{\hat{\theta}}(t|z)$ is an estimator of $S_T(t|z)$ based on the training data D. And by definition of (1.4), the prediction error $\pi_{3t}(M; F_T, \tilde{H}_Z)$ is,

$$\begin{aligned}\pi_{3t}(M; F_T, \tilde{H}_Z) &= \frac{1}{n}\sum_{i=1}^{n}[S_T(t|z_i)(1 - Pr(S_{\hat{\theta}}(t|z_i) > 0.5) \\ &\quad +(1 - S_T(t|z_i))Pr(S_{\hat{\theta}}(t|z_i) > 0.5)].\end{aligned} \tag{4.3}$$

If we use the estimator $S_{\hat{\theta}}(t|z)$ for $S_T(t|z)$, we obtain a model-based estimator of π_{2t},

$$\hat{\pi}_{2t}^m(M; F_T, z) = S_{\hat{\theta}}(t|z)\widehat{Pr}(S_{\hat{\theta}}(t|z) \leq 0.5) + (1 - S_{\hat{\theta}}(t|z))\widehat{Pr}(S_{\hat{\theta}}(t|z) > 0.5).$$

$Pr(S_{\hat{\theta}}(t|z) \leq 0.5)$ can be estimated by simulation, in order to allow for model selection in the estimation/model fitting process. Both parametric and nonparametric bootstrap procedures can be used to get the estimate $\widehat{Pr}(S_{\hat{\theta}}(t|z) \leq 0.5)$. In a parametric procedure, we simulate K sets $D_1^*, \ldots, D_K^*$ of training data $\{(Y_i^*, z_i), i = 1, \ldots, n\}$ using $F_{\hat{\theta}}(y|z) = 1 - S_{\hat{\theta}}(y|z)$ and apply the modeling procedure to D_k^* to produce the predictor $\widehat{W}_t^{*k} = I[S_{\hat{\theta}_k^*}(t|z) > 0.5]$, where $\hat{\theta}_k^*$ is based on D_k^*. Note that we also need to generate censoring times for each D_k^*; they are generated using the model $\hat{S}^c(c|z)$ fitted in connection with IPCW estimation described below. This then gives

$$\widehat{Pr}(S_{\hat{\theta}}(t|z) \leq 0.5) = \frac{1}{K}\sum_{k=1}^{K} I(S_{\hat{\theta}_k^*}(t|z) \leq 0.5), \tag{4.4}$$

for any z. We can alternatively draw K nonparametric bootstrap samples and apply the modeling procedure for each sample to get (4.4). This has the advantage that censoring times do not need to be generated, since we simply resample with replacement from $\{(t_i, \delta_i, z_i),\ i = 1, \ldots, n\}$ to get $D_1^*, \ldots, D_K^*$.

An estimator of π_t in (4.3) is then

$$\begin{aligned}\hat{\pi}_t^m &= \frac{1}{n}\sum_{i=1}^{n} \widehat{\pi}_{2t}^m(z_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} [S_{\hat{\theta}}(t|z_i)\widehat{Pr}(S_{\hat{\theta}}(t|z_i) \leq 0.5) \\ &\quad + (1 - S_{\hat{\theta}}(t|z_i))\widehat{Pr}(S_{\hat{\theta}}(t|z_i) > 0.5)]. \end{aligned} \tag{4.5}$$

The second approach in section 1.2.1 involves the apparent loss (AL), to which we can add an "optimism" adjustment Ω_t,

$$\hat{\pi}_t^A = \mathrm{AL}_t + \hat{\Omega}_t. \tag{4.6}$$

When Y is subject to censoring, W_t may not be observed, and we use IPCW with fitted model $\hat{S}^c(c|z)$ for the censoring time survivor function, as described in section 2.2.1. This gives the IPCW estimator of apparent loss,

$$\hat{\mathrm{AL}}_t \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{S}^c(t_i \wedge t|z_i)} I(w_{i,t} \neq \hat{w}_{i,t}) \tag{4.7}$$

where $w_{i,t} = I(y_i > t), \hat{w}_{i,t} = I(S_{\hat{\theta}}(t|z_i) > 0.5),\ \Delta_i$ indicates whether $I(w_{i,t} \neq w_{i,t})$ is observed, and $X \wedge Y = \min(X, Y)$.

The adjustment term proposed by Efron (2004) takes the form $\Omega_t = \frac{2}{n}\sum_{i=1}^{n}\mathrm{cov}(w_{i,t}, f(\hat{w}_{i,t}))$ (1.15). In the case of the 0-1 loss, $f(\hat{w}_t) = \hat{w}_t$. We use a model-based approach, as suggested by Efron (2004), to estimate $\mathrm{cov}(w_t, \hat{w}_t)$ by simulation from $\hat{F}_T$.

The V-fold cross-validation estimator $\hat{\pi}^{cv}$ replaces $\hat{w}_{i,t}$ in (4.7) with $\hat{w}_{i(-v),t}$, giving

$$\hat{\pi}_t^{cv} = \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v}\frac{\Delta_i}{\hat{S}^c(t_i \wedge \tau|z_i)} I(w_{i,t} \neq \hat{w}_{i(-v),t}). \tag{4.8}$$

Confidence interval estimates for π_t are obtained as described in section 2.4, using any of the three estimators of π_t. Simulation results for point estimates and confidence interval estimates are given in the next section.

4.2 Simulation Results

We used the simulation model (3.3) and the same modelling procedures described in chapter 3. For the binary 0-1 loss, we investigated the prediction error for survival status W_t at ten evenly spaced time points, $t_1 = 5,\ t_2 = 10, \ldots,\ t_{10} = 50$ months, with 500 simulation runs. In every simulation, survival status was predicted for subject i at the ten time points t_j, $\hat{W}_{i,t_j} = I(S_{\hat{\theta}}(t_j|z_i) > 0.5),\ j = 1, \ldots, 10$, where $S_{\hat{\theta}}(t_j|z_i)$ is given by the fitted model following the variable selection.

To obtain the model-based estimator $\hat{\pi}_t^m$ in (4.5), the quantity

$$\widehat{Pr}(S_{\hat{\theta}}(t|z) \leq 0.5)$$

was estimated using either parametric bootstrap or nonparametric bootstrap samples using (4.4). Parametric bootstrap refers to simulation from the fitted model $F_{\hat{\theta}}$ following variable selection. Nonparametric bootstrap refers to simple random sampling with replacement from D. We used $K = 100$ bootstrap samples for the estimates $\hat{\pi}_t^m$ and $B = 30$ samples for obtaining $\hat{\text{var}}(\hat{\pi}_t^m)$ (see Figure 2.2). A small B is used here for computational reasons; for each subject we predict W_t and estimate prediction error ten times. The pseudo training data at the second level of simulation, D_{bk}^*, was generated by either parametric or nonparametric bootstrap, according to the method used to estimate $\widehat{Pr}(S_{\hat{\theta}}(t|z) \leq 0.5)$ for $\hat{\pi}_t^m$. We note that the nonparametric bootstrap reduces the effective sample size, which increases the variability of the estimator. The second level estimator $\hat{\pi}_b^*$ is obtained by bootstrap sampling from D_b^*. When D_b^* is itself a nonparametric sample from D, the effective sample size for $\hat{\pi}_b^*$ is much smaller than the original sample size n of D.

Note that we can use a nonparametric approach to obtain the model-based estimator $\hat{\pi}_t^m$ for the 0-1 loss, but can not do so for the absolute error loss. This is because we need a training data set D^* and an *independent* test data set Y^* to estimate $\hat{\pi}^m$ for absolute error loss. If both Y^* and D^* are nonparametric bootstrap samples from the same data D, they are not independent anymore, and thus can not be used for estimating $\hat{\pi}^m$. For 0-1 loss, we only need D^* to estimate $\hat{\pi}_t^m$.

For the AL and CV approaches, we used the penalty-adjusted apparent loss $\hat{\pi}_t^A$ and 5-fold cross-validation $\hat{\pi}_t^{cv}$, respectively. The variance estimates for $\hat{\pi}_t$ were obtained with $B = 100$ bootstrap samples (both parametric based on $F_{\hat{\theta}}$ and nonparametric). We choose a small B because for each subject we need to predict W_t ten times. For single data sets, larger values can be used. Confidence intervals are based on treating $\hat{\pi}_t$ as normally distributed random variables; a bias correction for $\hat{\pi}_t^m$ is also considered and results are shown in Table 4.1. The percentile and basic percentile method were not considered here since $B = 100$ is too small to give accurate coverage probabilities.

Figure 4.1 shows the simulation results. The true π_t are given by lines, and the average point estimates of misclassification error at $t_1, \ldots, t_{10}$ are given by symbols. π_t is estimated using (4.3), i.e.

$$\begin{aligned}\pi_t &= \frac{1}{n}\sum_{i=1}^{n}[S_T(t|z_i)Pr(S_{\hat{\theta}}(t|z_i) \leq 0.5) \\ &\quad +(1 - S_T(t|z_i))Pr(S_{\hat{\theta}}(t|z_i) > 0.5)],\end{aligned}$$

where $S_T(t|z_i)$ is known and $Pr(S_{\hat{\theta}}(t|z_i) \leq 0.5)$ is estimated with 5000 samples generated under the true Weibull model (3.3), and accounts for the variable selection scheme under either the Weibull or Lognormal model. The true misclassification error is slightly smaller under the Weibull family as indicated by the lines corresponding to the Weibull model lying beneath the lines corresponding to the Lognormal model (see graphs in the right panels of Figure 4.1).

Figure 4.1 indicates that the adjusted apparent loss estimator $\hat{\pi}^A$ is almost unbiased, even when sample size is small ($n = 65$). The model-based penalty term $\hat{\Omega}$ performs well and corrects the bias of apparent loss. Similar to the results for absolute error loss, we find that $\hat{\pi}^{cv}$ overestimates π somewhat when $n = 65$ but is very accurate for $n = 400$. The model-based estimators $\hat{\pi}^m$ tends to underestimate π when $n = 65$. When $n = 400$, $\hat{\pi}^m$ for the Lognormal model, which corresponds to the upsidedown triangle in the bottom right graph, drifts away from the true error π_t as t increases. This result agrees with our discussion in section 2.3: $\hat{\pi}^m_t$ converges to $\pi_3(M; F_{\theta^*})$, not the true error of interest $\pi_3(M; F_T)$. When t increases from 35 to 50 months, the biases of the point estimates $\hat{\pi}^m$ also increase. As a result, the undercoverage of the π_t becomes more serious as t increases.

The coverage proportions of the ten time points are averaged and summarized in Table 4.1. The results in Table 4.1 agree with those in Table 3.3 in general: the model-based procedure performs well under the correct model with nonparametric bootstrap (with or without bias

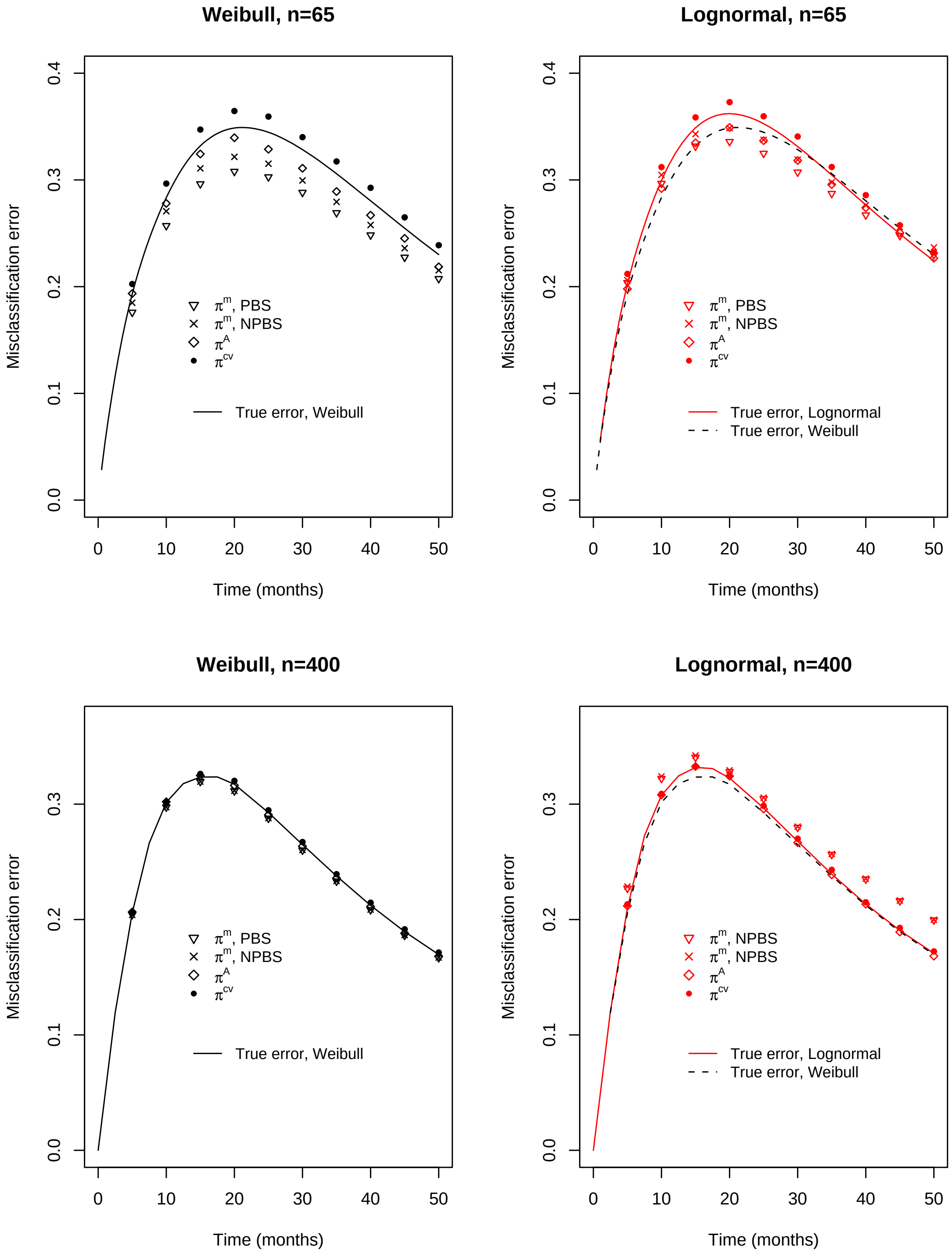

Figure 4.1: Simulation results for estimates of prediction error based on binary loss. The graphs in the left panels are for the Weibull model, and the graphs in the right panels are for the Lognormal model. The top panels correspond to sample size 65 and the bottom panels correspond to sample size 400. The lines give the true prediction error estimated from 5000 simulations, and the points give the $\hat{\pi}$ averaged over 500 simulation runs.

Model fitted	Estimator	PBS	NPBS
		NA/Bias corrected NA	NA/Bias corrected NA
Weibull $n = 65$	Model-based $\hat{\pi}^m$	0.758/0.857	0.932/0.927
	Apparent Loss (adj.) $\hat{\pi}^A$	0.921	0.956
	5 fold CV $\hat{\pi}^{cv}$	0.946	0.966
Lognormal $n = 65$	Model-based $\hat{\pi}^m$	0.879/0.858	0.920/0.869
	Apparent Loss (adj.) $\hat{\pi}^A$	0.927	0.957
	5 fold CV $\hat{\pi}^{cv}$	0.944	0.962
Weibull $n = 400$	Model-based $\hat{\pi}^m$	0.924/0.934	0.950/0.951
	Apparent Loss (adj.) $\hat{\pi}^A$	0.942	0.963
	5 fold CV $\hat{\pi}^{cv}$	0.950	0.967
Lognormal $n = 400$	Model-based $\hat{\pi}^m$	0.766/0.675	0.762/0.672
	Apparent Loss (adj.) $\hat{\pi}^A$	0.952	0.967
	5 fold CV $\hat{\pi}^{cv}$	0.960	0.969

Table 4.1: Average coverage proportions of the ten time points for binary 0-1 loss, at nominal 0.95 confidence levels with different methods based on 500 simulations. PBS and NPBS refer to parametric and nonparametric bootstrap, respectively, as described in section 2.4. NA stands for normal approximation. Model-based estimates are subject to bias correction, and coverage proportions for both uncorrected and corrected point estimates are given. The adjusted apparent loss estimates and 5-fold cross-validation estimates are not corrected for bias, hence only one value is given.

correction); confidence intervals based on the adjusted apparent loss estimator $\hat{\pi}^A$ perform well under both models except for the slight undercoverage when parametric bootstrap is used in the small sample case; confidence intervals based on $\hat{\pi}^{cv}$ without bias correction perform well under all circumstances investigated.

Figure 4.2 shows the coverage proportions of the individual time points for nominal 0.95 confidence intervals based on the 500 simulations. Different symbols are used for different methods of obtaining confidence intervals. Examination of Figure 4.2 indicates that

1. Intervals based on $\hat{\pi}^m$ perform well under the Weibull model with bias correction and nonparametric bootstrap when sample size is 65. When sample size is 400, both parametric and nonparametric bootstrap perform well with bias correction.

2. Intervals based on $\hat{\pi}^A$ perform well under the Weibull model. The Lognormal model gives less satisfactory coverage probabilities under the same settings, though the coverage proportion is good when averaged over the ten time points (Table 4.1) .

3. Intervals based on $\hat{\pi}^{cv}$ work well when parametric bootstrap is used under the correct model. Some mild overcoverage is observed for intervals based on $\hat{\pi}^{cv}$ in other settings.

To summarize, parametric model-based methods are susceptible to model misspecification. Presumably this is due to the same reasons discussed in section 3.5 for absolute error loss. Overall, cross-validation methods work well in conjunction with nonparametric bootstrap resampling, and for reasonably large samples, methods based on apparent error and nonparametric bootstrap resampling also work well.

Molinaro et al. (2005) has investigated the estimation of prediction error for the binary classification problem with variable (feature) selection and a number of parametric and nonparametric models. They conducted an extensive simulation study to compare estimators based on v-fold cross-validation, leave-one-out cross-validation, Monte Carlo cross-validation,

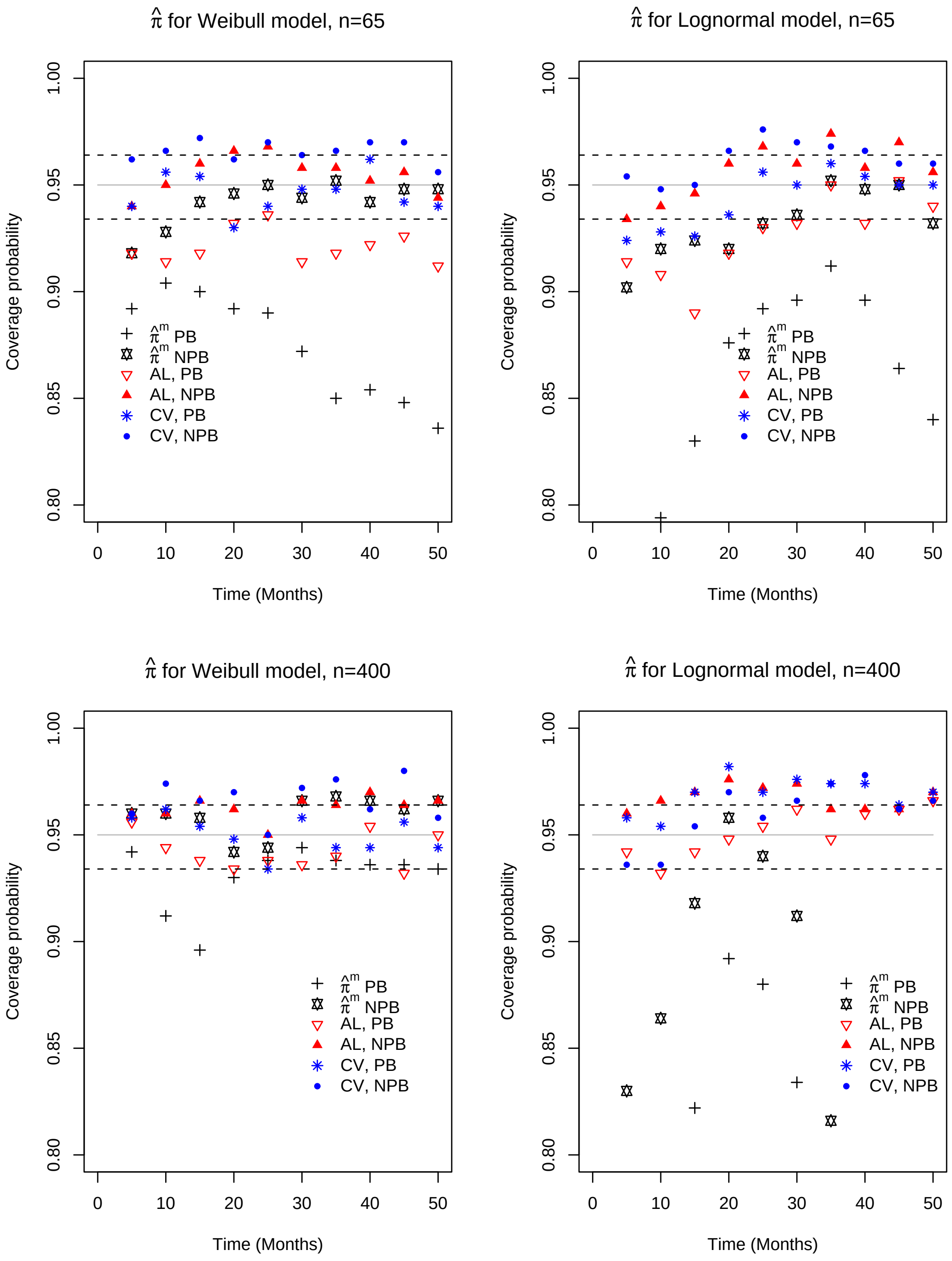

Figure 4.2: The coverage proportions of the ten time points for binary 0-1 loss, at nominal 0.95 confidence levels with different methods based on 500 simulations. The graphs in the top panels are for $n = 65$, and the graphs in the bottom panels correspond to $n = 400$. The left panels are for the Weibull model and the right panels are for the Lognormal model. In each graph, the solid line is the reference line for the nominal 95% level and the dashed lines give the pointwise 0.95 probability intervals $[p_L, p_U]$ based on 500 Bernoulli trials with $p = 0.95$.

and 0.632+ bootstrap. Tibshirani and Knight (1999) studied model and variable selection and its connection with prediction error in general settings. Both papers suggest that cross-validation tends to perform well as far as the point estimation of prediction error is concerned. We have studied prediction error estimation using absolute error and binary loss functions for censored survival data. Our results agree with their findings and furthermore we show that confidence intervals based on the CV estimator with the nonparametric bootstrap procedure perform well.

Chapter 5

Applications

In this chapter, we consider prediction error estimation for two data sets, (i) the multiple myeloma data on which the preceding simulation studies were based, and (ii) the Mayo Clinic Primary Biliary Cirrhosis (PBC) data, which has been considered by Gerds and Schumacher (2006) and other authors.

5.1 Multiple Myeloma Data

The multiple myeloma data (Krall et al. 1975) is analyzed with the 9 covariates described in Table 3.1. Weibull and Lognormal models are fitted with variable selection, which correspond to using model (3.3) and assuming ε_i is extreme value and normally distributed, respectively. The reduced Weibull model has three variables: log(Blood Urea Nitrogen), Hemoglobin and Serum calcium. The reduced Lognormal model has four variables: log(Blood Urea Nitrogen), Hemoglobin, Serum Calcium and Fractures. Residual probability plots suggest that both reduced models provide adequate fits to the data (Figure 5.1). Besides the Weibull and Lognormal model, we also fit a semiparametric Cox PH model with same variable selection criterion. The

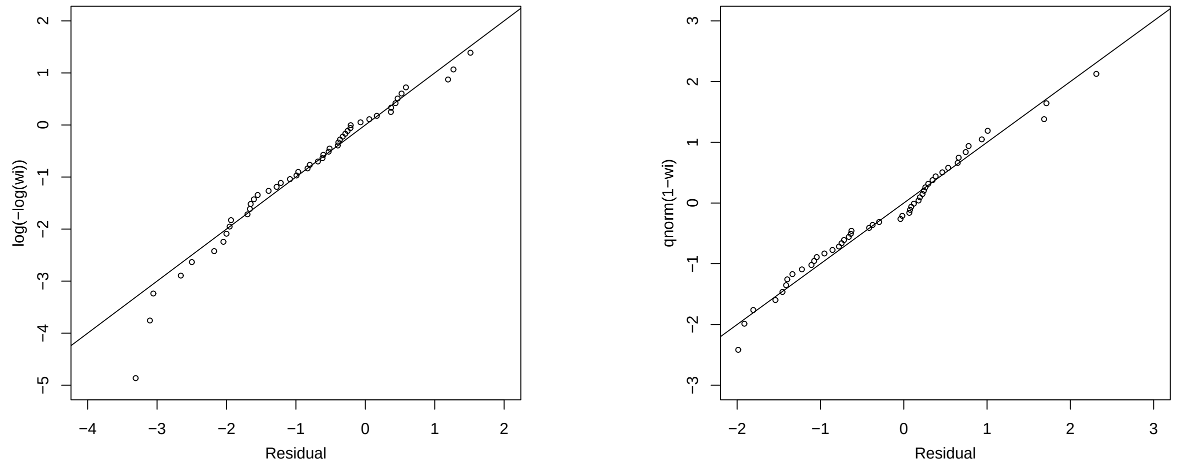

Figure 5.1: Residual plots for the reduced Weibull (left panel) and Lognormal (right panel) models for Multiple Myeloma data.

reduced Cox PH model has two variables, log(Blood Urea Nitrogen) and Hemoglobin.

5.1.1 Absolute error loss

We are interested in estimating the prediction error associated with these three families of models with the variable selection procedure. Based on the simulation results of chapter 3, the parametric bootstrap procedure for estimation of var($\hat{\pi}^A$) and var($\hat{\pi}^{cv}$) are not pursued. For the model-based point estimates $\hat{\pi}^m$ and var($\hat{\pi}^m$), we used $K = B = 500$ bootstrap samples. We also used $B = 500$ nonparametric bootstrap samples for the estimation of confidence intervals using AL ($\hat{\pi}^A$) and 5-fold CV estimates ($\hat{\pi}^{cv}$), and $B = 100$ bootstrap samples for the more computationally demanding leave-one-out CV (LOOCV), which is 65-fold. Confidence intervals are based on the approximate normality of log($\hat{\pi}$) since the sample size is small, and bias correction was used for $\hat{\pi}^m$ and $\hat{\pi}^A$.

In order to use IPCW, we modeled the censoring times for an estimate of $S^c(c|z)$ (2.7).

Note that the covariates used for modeling the censoring times are those in the reduced final model for survival times, because the IPCW method is valid provided $Y \perp C$ given a set of covariates Z. Suppose survival time Y depends on Z_1 and censoring time C depends on Z_1 and Z_2, it can be shown that Y is independent of C given only Z_1:

$$\begin{aligned} f(y, c|z_1) &= \int f_0(y, c, z_2|z_1) dz_2 \\ &= \int f_1(y, c|z_1, z_2) g(z_2|z_1) dz_2 \\ &= \int f_Y(y|z_1, z_2) f_C(c|z_1, z_2) g(z_2|z_1) dz_2 \\ &= f_Y(y|z_1) \int f_C(c|z_1, z_2) g(z_2|z_1) dz_2 \\ &= f_Y(y|z_1) f_C(c|z_1). \end{aligned}$$

Therefore, the set of covariates C is regressed on can be a subset of the covariates in the survival model for Y. For the example above, the IPC weight α is estimated by $\hat{S}^c(c|z_1)$.

For the multiple myeloma data, we found that the censoring times are roughly independent of the covariates used in the survival models, such as log(Blood Urea Nitrogen), Hemoglobin, Serum Calcium and Fractures (all P-values > 0.05). Hence, a random censoring mechanism was assumed and the Kaplan-Meier estimate $\hat{S}^c(c)$ was used for the censoring survivor function, which was also used to generate the censoring times for parametric bootstrap samples.

The point estimates of the prediction error (1.4) with absolute relative error loss truncated at $\tau = 50$ months and the corresponding confidence intervals for the three fitted models are reported in Table 5.1. For the Cox PH model, the estimated survival probability for each individual is given by

$$\hat{S}(y|z_i) = \hat{S}_0(y)^{\exp(z_i^T \hat{\beta})}, \tag{5.1}$$

where $\hat{S}_0(y)$ is the estimated baseline survival probability. Then the predicted median survival time for y_i is

$$\hat{y}_i = \underset{t}{\operatorname{argmax}} \{\hat{S}(t|z_i) > 0.5\},$$

	Weibull		Lognormal		Cox PH	
	$\hat{\pi}$	95% CI	$\hat{\pi}$	95% CI	$\hat{\pi}$	95% CI
$\hat{\pi}^m$ PBS	0.98	(0.76, 1.30)	0.87	(0.71, 1.10)	0.96	(0.76, 1.23)
$\hat{\pi}^m$ NPBS	1.04	(0.81, 1.42)	0.87	(0.71, 1.10)	0.97	(0.73, 1.34)
$\hat{\pi}^A$ NPBS	0.89	(0.69, 1.19)	0.88	(0.69, 1.16)	0.86	(0.55, 1.40)
$\hat{\pi}^{cv}$ NPBS	0.91	(0.68, 1.23)	0.91	(0.68, 1.21)	0.91	(0.67, 1.25)
LOOCV NPBS	0.94	(0.70, 1.28)	0.95	(0.72, 1.27)	0.93	(0.68, 1.27)

Table 5.1: The estimated prediction error and 95% confidence interval under the Weibull, Lognormal and Cox PH models with absolute error loss function truncated at $\tau = 50$ months. $\hat{\pi}^m$ is the model-based estimator, $\hat{\pi}^A$ refers to the AL method, $\hat{\pi}^{cv}$ to the 5-fold CV method, and LOOCV stands for the leave-one-out CV method. The confidence interval is based on the approximation that $\log(\hat{\pi})$ is normally distributed. The point estimates $\hat{\pi}^m$ and $\hat{\pi}^A$ reported here are corrected for bias with the estimates from bootstrap samples, as described in section 2.4.

where t is the set of distinctive failure times in the training data D. It is possible that $S(t_{max}|z_i) > 0.5$ for some z_i where t_{max} denote the maximum observed failure time in D. In this case the predicted median survival time $\hat{y}_i > t_{max}$ and is not available. However, since we consider the truncated loss function and choose $\tau \leq t_{max}$, the predictor is given by $\min(\hat{Y}, \tau) = \tau$ if $\hat{Y}$ is greater than t_{max}. To obtain the model-based estimator $\hat{\pi}^m$, we take parametric bootstrap samples from the Cox PH model by treating the step function $\hat{S}(y|z_i)$ in (5.1) as the survivor function of a multinomial random variable Y_i, and we generate survival time Y_i by sampling this distribution. Therefore the possible values of Y_i are limited to the observed failure times in D.

The point estimates are similar under different approaches and the three model families, except for the estimates based on $\hat{\pi}^m$ under the Weibull and Cox PH model: they are slightly greater than the rest of the estimates. The model-based approach is not very reliable as discussed in section 3.5. It is included merely for comparison and comprehensiveness. In addition, we note that for Weibull and Lognormal models, the confidence intervals based on the CV estimates are slightly wider than those of the corresponding AL estimates, and confidence intervals based on the Lognormal model are slightly narrower than those of the Weibull and Cox PH models. The confidence intervals under the Cox PH model are comparable to those of the Weibull models, except for the one based on the AL estimate, which is the widest among all estimators and model families considered.

We also estimate the prediction error of the marginal model as define in section 1.2.2 with the IPCW approach. The marginal model in survival settings is given by the Kaplan-Meier estimate $\hat{S}(y)$. The 5-fold cross-validation estimator of $\hat{\pi}_0^{cv}$ (1.22) is 1.001 and the 95% confidence interval is (0.75, 1.25), which are in close agreement with the estimates given in Table 5.1. For the multiple myeloma data, the covariates and the survival models seem to have little if any predictive power.

5.1.2 Binary loss

For the Weibull and Lognormal model families and modeling procedures described above, we also estimated the binary misclassification loss and confidence intervals for the prediction of $W_t = I(Y > t)$ at $t = 10,\ 20,\ 30,\ 40,\ 50$ months with the approaches described in chapter 4. The point estimates are plotted in Figure 5.2. The 95% confidence intervals for month 20 and 40 are shown in the same figure for illustration. For comparison, we also plot in the same graph the misclassification error of the marginal model (5-fold CV estimate).

Figure 5.2 shows that the model-based estimates agree well between parametric and non-parametric procedures within each model family. Their confidence intervals are narrower than the confidence intervals of other estimates. As previously reported for absolute error loss, the confidence intervals based on CV estimates are slightly wider than the intervals based on adjusted apparent error estimates. For most time points, the confidence intervals given by the

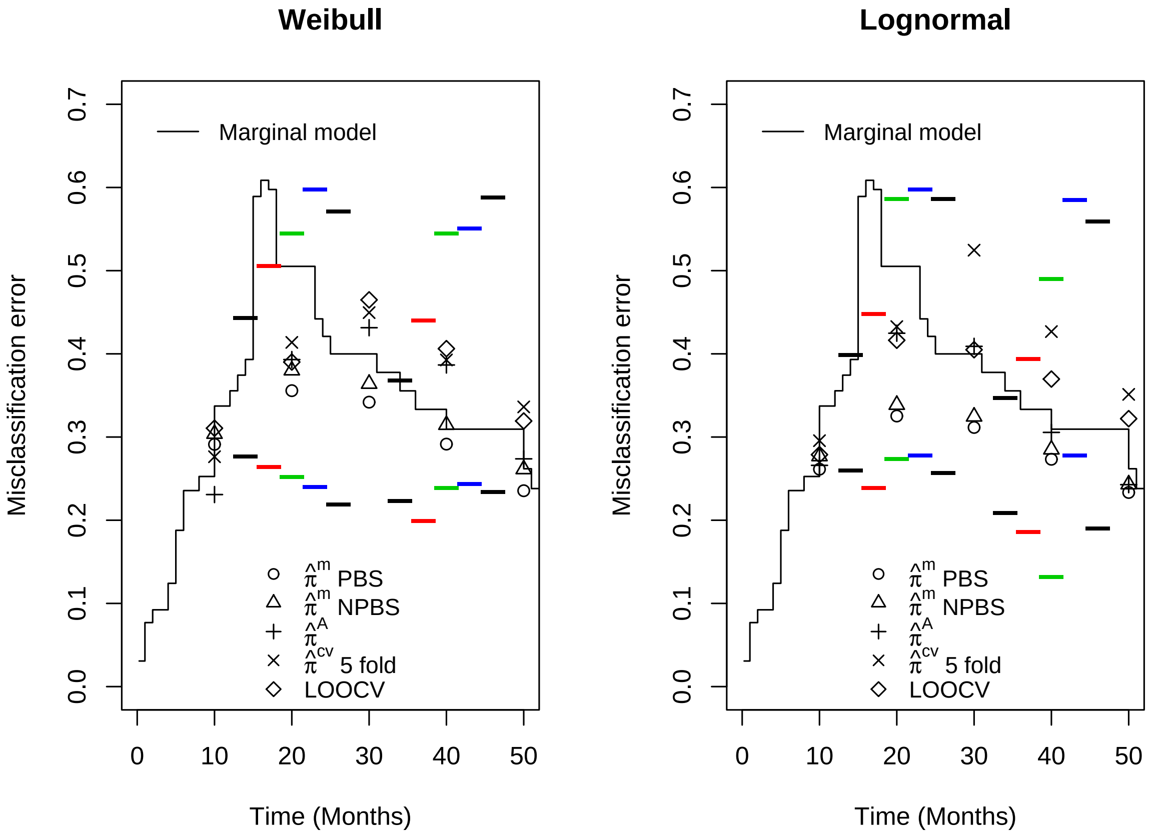

Figure 5.2: Estimates of prediction error based on binary loss at $t = 10, 20, 30, 40, 50$ months. The graph on the left is for the Weibull model, and the graph on the right is for the Lognormal model. The 95% intervals for the misclassification error are given by pairs of "-" symbols for month 20 and 40. For each time point, the five intervals from left to right correspond to the five estimators from top to bottom (see legend). The solid line shows the misclassification error (5-fold CV estimate) of the marginal model.

regression model contain the corresponding estimated misclassification error of the marginal model. It shows that not much prediction power is gained when the regression model is used. This finding agrees with the results obtained for absolute error loss, i.e. the prediction error estimates are comparable between the regression models and the marginal model.

5.1.3 Estimated ROC curve

The misclassification error rate is one way of assessing the classification rules. As discussed in section 2.1.3, an ROC curve is often used to evaluate the classifier performance for its ranking ability. We estimate the ROC curve with the IPCW approach described in section 2.2.1 for the multiple myeloma data. For illustration, we choose $t_0 = 10$ months, at which time six out of 65 subjects are censored. Among the remaining 59 patients, twenty-one have failed. Our objective is to identify these early failures using the survival models. Thus we define $W_{i,t_0} = I(Y_i \leq t_0)$, for which the predicted class is given by $\hat{W}_{i,t_0} = I(S_{\hat{\theta}(-i)}(t_0|z_i) \leq c)$, and $\hat{\theta}(-i)$ is a LOOCV estimate. In this case, the ROC curve consists of points with (FPR(c), TPR(c)) as their X-Y coordinates, moving from (0, 0) to (1, 1) as c increases from [0,1].

Figure 5.3 displays the estimated ROC curve of the reduced Weibull model for $t_0 = 10$ months, assuming random censoring. As we can see, it is relatively close to the reference line for a random ranking model, indicating the lack of ranking power of the survival model for detecting early failures.

5.2 Mayo PBC Data

The second example is based on the well-known Mayo primary biliary cirrhosis data (Therneau and Grambsch 2000, appendix D.2). The data set is available in R library "survival" and contains 418 subjects and 18 prognostic variables. There are missing covariate values for some

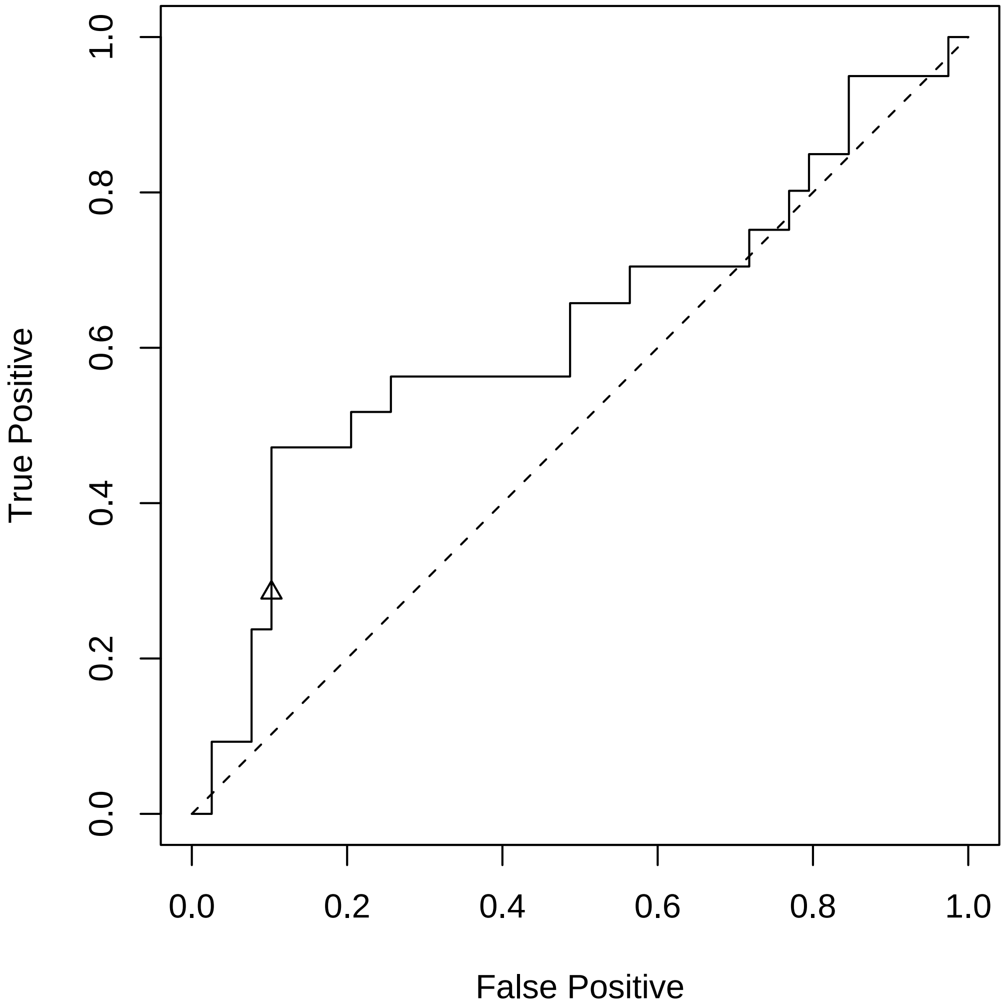

Figure 5.3: ROC curve of the reduced Weibull model for identifying the subjects who died within 10 months after diagnosis. △ displays (FPR(0.5), TPR(0.5)) based on the classification rule $\hat{W}_{i,t_0} = I(S_{\hat{\theta}(-i)}(t_0|z_i) \leq 0.5)$.

subjects, especially the 106 subjects who did not participate in the associated randomized clinical trial. We analyze the data from 276 subjects that have complete baseline covariate information. Others (e.g. van Houwelingen and le Cessie 1990; Gerds and Schumacher 2006) have analyzed this data set using five significant covariates suggested by Fleming and Harrington (1991). These five variables have no missing values and their data analysis are apparently based on all 418 subjects. Since binary variable "edema" indicates the presence and absence of edema and for variable "edtrt", 0 = no edema, 0.5 = edema present but not treated or successfully treated, 1 = edema present and did not respond to treatment, the information contained in "edema" is in "edtrt". Thus we omit "edema" from the covariates. The same 276 subjects and 17 variables were studied by Tibshirani (1997) and Soh et al. (2008). Below is a list of the variables included in our data analysis.

Y: Survival time.

δ: 1 if Y is time to death, 0 if time to censoring.

X_1: Treatment code, 1 = D-pencillamine, 2 = placebo.

X_2: Age in years.

X_3: Sex, 0 = male, 1 = female.

X_4: Presence of ascites, 0 = no, 1 = yes.

X_5: Presence of hepatomegaly, 0 = no, 1 = yes.

X_6: Presence of spiders, 0 = no, 1 = yes.

X_7: Presence of edema, 0 = no, 0.5 = yes but either not treated or responded to diuretic treatment, 1 = yes, did not respond to treatment.

X_8: Serum bilirubin, in mg/dl.

X_9: Serum cholesterol, in mg/dl.

X_{10}: Albumin, in g/dl.

X_{11}: Urine copper, in kg/day.

X_{12}: Alkaline phosphatase, in U/litre.

X_{13}: SGOT, a liver enzyme, in U/ml.

X_{14}: Triglycerides, in mg/dl.

X_{15}: Platelet count; coded value is number of platelets per cubic ml of blood divided by 1000.

X_{16}: Prothrombine time, standardized blood clotting time, in seconds.

X_{17}: Histologic state of disease, graded 1, 2, 3 or 4.

Continuous covariates are log transformed following preliminary checks of the data and suggestions from data analysis by Fleming and Harrington (1991). Sixty percent of the survival times are censored and the maximum follow-up time is 13 years. We therefore choose $\tau = 10$ years as the upper bound for the absolute error loss. The marginal survival and censoring probabilities at τ are 0.42 and 0.23, respectively.

5.2.1 Absolute error loss

Weibull, Lognormal and Cox PH models were fitted with variable selection. The reduced Weibull model has eight variables and they are: X_2, X_7, X_8, X_{10}, X_{11}, X_{13}, X_{16}, X_{17}. The reduced Lognormal model also includes eight variables, seven of which are used by the Weibull model and the other variable is X_4, which replaces X_{10} in the above list. There are seven variables in the reduced Cox PH model, all of which are in the reduced Weibull model. X_{13} is the variable in the Weibull model but not in the reduced Cox PH model. The residual probability plots suggest that the Weibull model provides a slightly better fit than the Lognormal model (Figure 5.4).

	Random Censoring						Independent Censoring					
	Weibull		Lognormal		Cox PH		Weibull		Lognormal		Cox PH	
	$\hat{\pi}$	95% CI	$\hat{\pi}$	95% CI	$\hat{\pi}$	95% CI	$\hat{\pi}$	95% CI	$\hat{\pi}$	95% CI	$\hat{\pi}$	95% CI
$\hat{\pi}^m$ PBS	0.39	(0.32, 0.45)	0.38	(0.34, 0.43)	0.38	(0.30, 0.46)	0.38	(0.32, 0.45)	0.38	(0.33, 0.43)	0.39	(0.31, 0.47)
$\hat{\pi}^m$ NPBS	0.40	(0.33, 0.47)	0.37	(0.31, 0.44)	0.41	(0.32, 0.51)	0.39	(0.32, 0.46)	0.37	(0.31, 0.44)	0.42	(0.33, 0.51)
$\hat{\pi}^A$	0.37	(0.30, 0.45)	0.41	(0.32, 0.49)	0.37	(0.27, 0.48)	0.33	(0.25, 0.42)	0.32	(0.23, 0.40)	0.35	(0.25, 0.46)
$\hat{\pi}^{cv}$	0.37	(0.28, 0.45)	0.40	(0.31, 0.49)	0.37	(0.28, 0.46)	0.35	(0.28, 0.46)	0.37	(0.28, 0.46)	0.35	(0.26, 0.45)
LOOCV	0.38	(0.29, 0.47)	0.41	(0.31, 0.51)	0.39	(0.29, 0.48)	0.36	(0.27, 0.45)	0.38	(0.29, 0.47)	0.38	(0.28, 0.47)

Table 5.2: The estimated prediction error and 95% confidence interval for the primary biliary cirrhosis data under the Weibull, Lognormal and Cox PH models, with absolute error loss function truncated at $\tau = 10$ years. $\hat{\pi}^A$ refers to the AL method, $\hat{\pi}^{cv}$ to the 5-fold CV method, and LOOCV stands for the leave-one-out CV method. The point estimates $\hat{\pi}^m$ and $\hat{\pi}^A$ are corrected for bias with the estimates from bootstrap samples, as described in section 2.4.

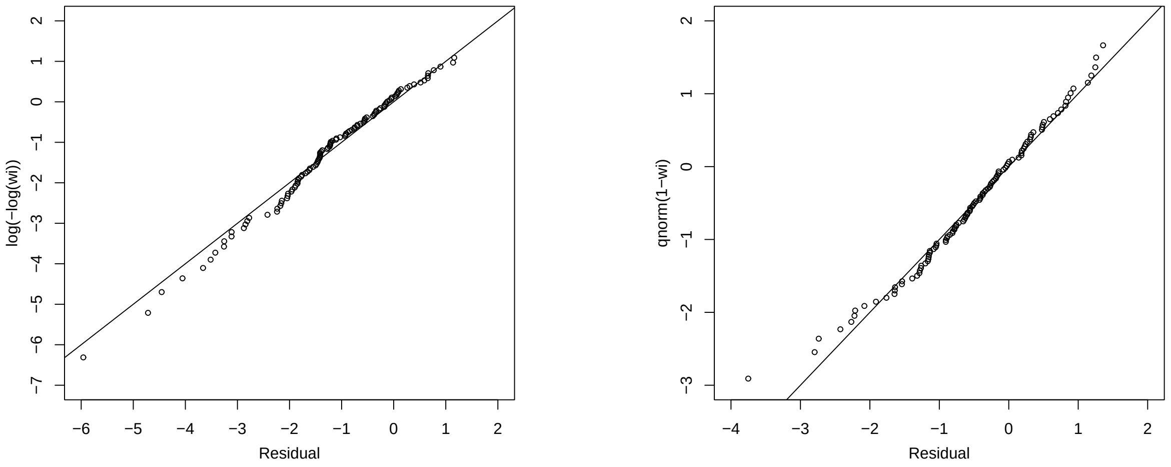

Figure 5.4: Residual plots for the reduced Weibull (left panel) and Lognormal (right panel) models for PBC data.

We estimated the prediction error associated with these families of models with the variable selection procedure. The point estimates of the prediction error with absolute log relative error loss truncated at $\tau = 10$ years and the corresponding confidence intervals for two IPC weights are reported in Table 5.2. One IPC weight is based on a random censoring assumption, so $\hat{S}^c(c|z)$ is simply the Kaplan-Meier estimate $\hat{S}^c(c)$ as in section 5.1.1, and the other weight is based on the independent censoring assumption where the censoring distribution is modeled with selected covariates. Weibull models provide adequate description of the censoring time distribution. Covariates were selected from the set used for the Weibull, Lognormal or Cox PH models for survival time, respectively. Variables significant at the 0.05 level for the Weibull, Lognormal and Cox PH cases are $(X_2,\ X_{10},\ X_{13},\ X_{16},\ X_{17})$, $(X_2,\ X_{13},\ X_{16},\ X_{17})$, and $(X_{10},\ X_{16},\ X_{17})$, respectively. We used $B = K = 500$ bootstrap samples for estimation of $\hat{\pi}^m$ and variances of $\hat{\pi}^m$, $\hat{\pi}^A$ and $\hat{\pi}^{cv}$. For the leave-one-out (276-fold) cross-validation, we used $B = 100$ bootstrap samples. Confidence intervals are based on the approximate normality of $\hat{\pi}$, and bias correction method was used for $\hat{\pi}^m$ and $\hat{\pi}^A$.

The point estimates of prediction error and the confidence intervals are similar across different models, regardless of the estimator used (Table 5.2). As expected, the model-based estimates for prediction error and confidence intervals are almost unchanged between the two censoring models. This is because no IPCW weights are used in model-based estimators. The independent censoring model gives slightly smaller point estimates and lower confidence limits for $\hat{\pi}^A$ and $\hat{\pi}^{cv}$. The prediction error of the null model was assessed with 5-fold CV, giving point estimate 0.56 and 95% confidence interval (0.47, 0.67). The estimates associated with the null model are greater than those of the regression models, indicating that the covariates and the survival models have some predictive power for this data set.

5.2.2 Binary loss

For the Weibull and Lognormal model families and modeling procedures described in 5.2.1, we estimated the misclassification error and confidence intervals for prediction of $W_t = I(Y > t)$ at $t = 1, 2, \ldots, 10$ years. The point estimates at the ten time points and the 95% confidence intervals for years 2, 5 and 9 are shown for the Weibull models in the left panels and for the Lognormal models in the right panels of Figure 5.5. The misclassification rates are similar between models, with rates slightly lower in the Weibull model. The width of the confidence intervals based on $\hat{\pi}^m$ is the smallest among the three estimators, and the width is consistent for time $=$ 2, 5 and 9 years. Meanwhile, the confidence intervals based on $\hat{\pi}^A$ and $\hat{\pi}^{cv}$ are comparable in width and they become wider as time increases from 2 to 9 years.

Figure 5.6 contrasts the misclassification error of the marginal model with that of a Weibull model under the random censoring assumption (other regression models give similar plots). We find that the error of the marginal model is above the 95% confidence intervals of the regression model except for the initial time period. It shows that the regression model can improve the prediction of survival status on PBC data, which agrees with our conclusion for the absolute error loss in section 5.2.1.

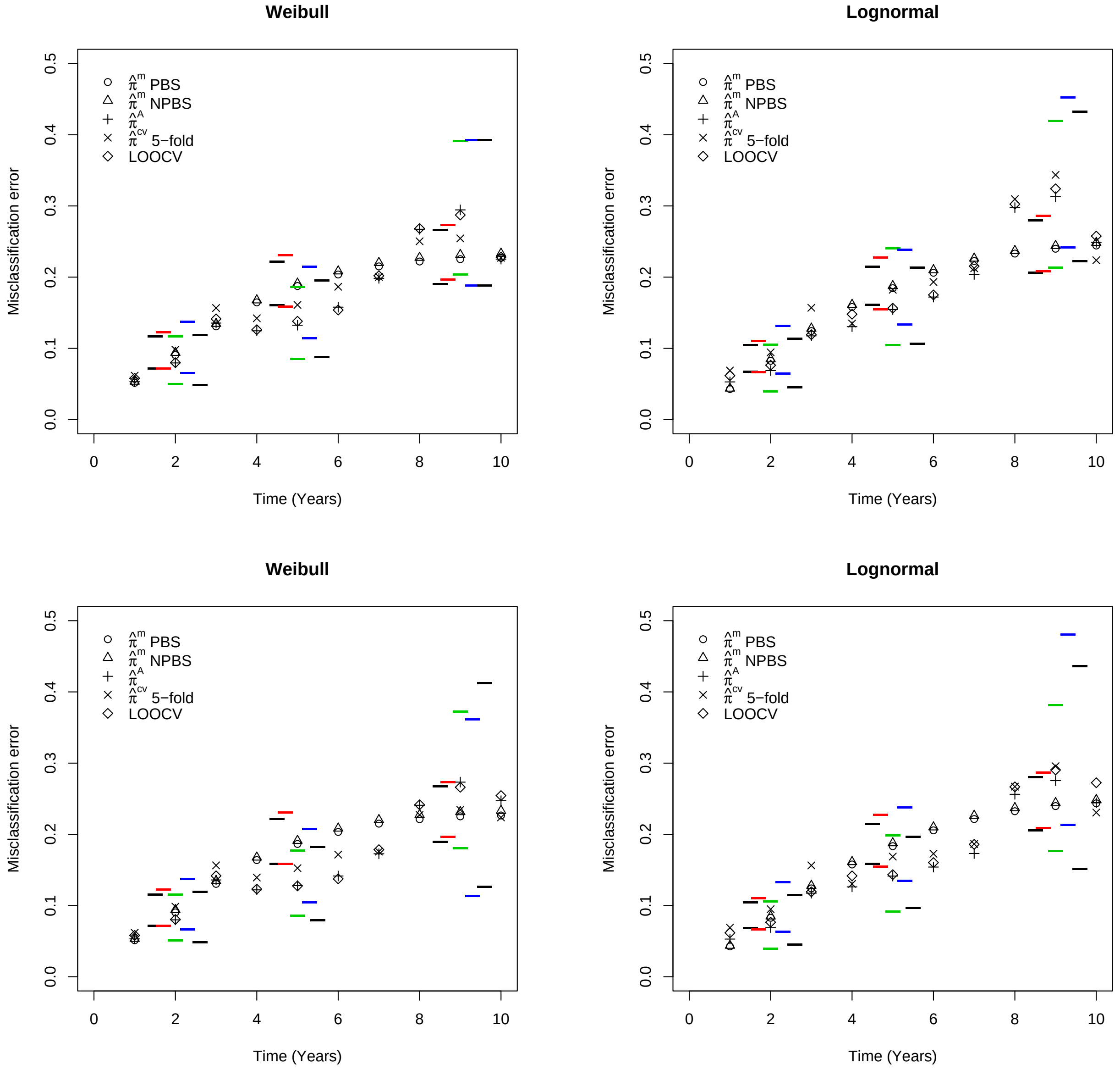

Figure 5.5: Estimates of prediction error based on binary loss at $t = 1, 2, \ldots, 10$ years. The graphs in the top panels assume random censoring and the graphs in the bottom panels assume independent censoring. The graphs in the left panels are for Weibull model, and the graphs in the right panels are for Lognormal model. The 95% intervals for the misclassification error are given by pairs of "-" symbols for $t = 2, 5$ and 9 years. For each time point, the five intervals from left to right correspond to the five estimators from top to bottom (see legend).

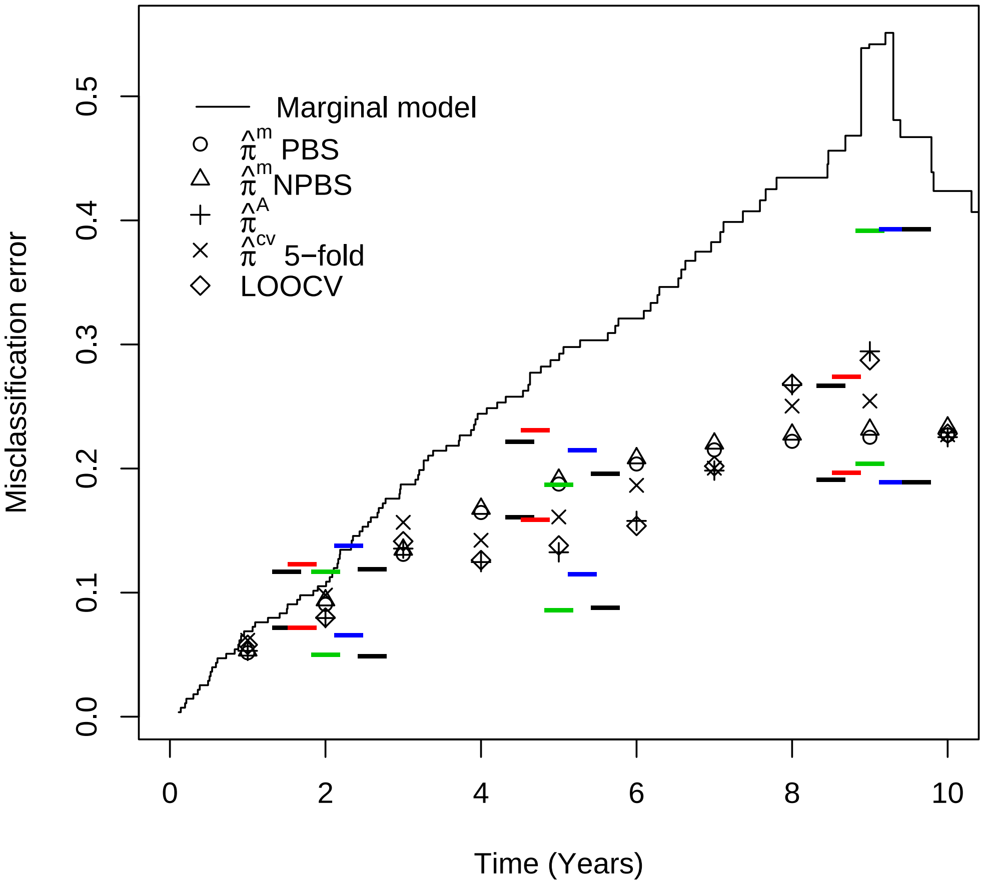

Figure 5.6: The estimates of misclassification error based on binary loss of the marginal model, in comparison with the point and interval estimates of the Weibull regression model under the random censoring assumption. The marginal error is estimated with 5-fold CV.

5.2.3 Estimated ROC curve

Figure 5.7 displays the estimated ROC curves for the Weibull model with random censoring in left panels and independent censoring in the right panels at $t = 2$ & 10 years. Subjects are ranked according to $S_{\hat{\theta}(-i)}(t|z_i)$, the LOOCV estimates of survival probability at t. Note that for $t_1 = 2$ and $t_2 = 10$ years, we define $W_{t_1} = I(Y \leq t_1)$ and $W_{t_2} = I(Y > t_2)$, respectively. This is because at $t_1 = 2$ years, our objective is to identify the early failures, but at $t_2 = 10$ years, our objective is to identify the long-term survivors.

All estimated ROC curves in Figure 5.7 are well above the reference line of an uninformative ranking model, suggesting good ranking ability of the Weibull model for identifying the early failures as well as the long-term survivors. The curves in the left and right panels are similar, indicating that the IPCW estimates of TPR(c) and FPR(c) are close under either random or independent censoring assumption. In addition, the curves in the top panels seem to be closer to the top and left borders of the unit square than the curves in the bottom panels do, implying that the discriminating power of the regression model may be larger for early failures than for long-term survivors.

At $t_1 = 2$ years, only one subject is censored. Out of the remaining 275 subjects, 29 subjects have failed. If we take $c = 0.5$, i.e. $\hat{W}_{i,t_1} = I(S_{\hat{\theta}(-i)}(t_1|z_i) \leq 0.5)$, we would predict 21 failures, of which 14 would be correctly classified. The overall misclassification rate in this case is 0.08. This classification rule is indicated by $\triangle$ in the top panels of Figure 5.7. When our primary objective is to identify the earlier failures, we may be willing to allow for a higher false positive rate. For example, if we take $c = 0.83$, i.e. $\hat{W}_{i,t_1} = I(S_{\hat{\theta}(-i)}(t_1|z_i) \leq 0.83)$, we would predict 55 failures, of which 24 would be correctly classified. This classification rule corresponds to $\times$ on the ROC curves (top panels, Figure 5.7). The overall misclassification error in this case is 0.13.

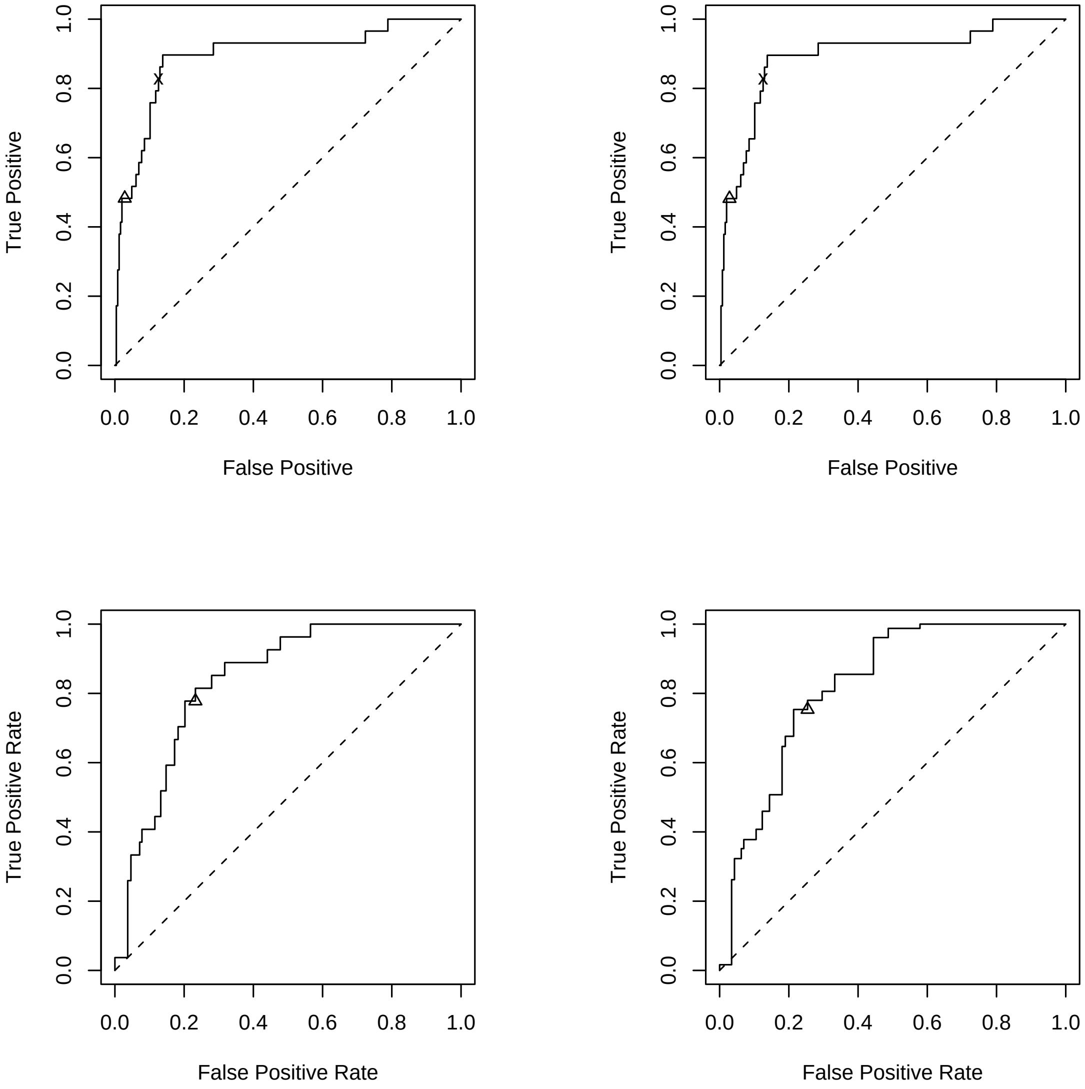

Figure 5.7: ROC curve for ranking ability of the Weibull regression model. The ROC curves in the left and right panels assume random and independent censoring, respectively. The top panels correspond to $t_1 = 2$ years and assess the ability of the regression model for identifying the early failures. The bottom panels correspond to $t_2 = 10$ years and assess the ability for identifying the long-term survivors. $\triangle$ displays (FPR(0.5), TPR(0.5)) based on the classification rules $\hat{W}_{i,t_1} = I(S_{\hat{\theta}(-i)}(t_1|z_i) \leq 0.5)$ and $\hat{W}_{i,t_2} = I(S_{\hat{\theta}(-i)}(t_2|z_i) > 0.5)$, for the top and bottom panels, respectively.

Chapter 6

Probabilistic Prediction of Survival Times

The prediction error and prediction power discussed in previous chapters focus on point prediction. As mentioned in chapter 1, there is another type of prediction, probabilistic prediction, which gives probabilities or prediction intervals for random variable Y. To do this, we use a probabilistic predictor $\hat{F}_p(y|z) = \widehat{Pr}(Y \leq y \mid Z = z)$. Since it specifies a distribution for Y given Z, the probabilistic predictor is also termed a predictive distribution. These two terms are used interchangeably henceforth. Examples of probabilistic prediction in survival contexts are easy to find. For example, clinicians may predict the 1-year survival probability to be 0.5 for one patient, and 0.1 for another. Furthermore, when clinicians make point predictions of the survival time, there is usually an unstated probability attached to that prediction. In this section, we consider methods of assessing and comparing predictive distributions for survival times.

Predictive distributions have received much attention, ranging from classical settings that do not involve covariates (e.g. Dawid 1984, Geisser 1993, Lawless and Fredette 2005) to complex settings involving meteorological processes (Gneiting et al. 2007). Several authors have studied probabilistic prediction for survival times and made important contributions, e.g. van Houwelingen and le Cessie (1990), Graf et al. (1999), Gerds and Schumacher (2006 & 2007). Nonetheless, there remain interesting and challenging issues. We address some of them, which include links between point and probabilistic predictors, and evaluation of the performance of a probabilistic predictor. Special attention is paid to the characteristics of survival data, such as long time span of studies and censoring of survival times.

A probabilistic predictor based on data $D = \{(y_i, z_i), i = 1, \ldots, n\}$ is denoted $\hat{F}_p(y|z)$, where z is a known vector of covariates. Often, $\hat{F}_p(y|z)$ is of the form $F_{\hat{\theta}}(y|z)$ where F_θ is a family of distributions indexed by the parameter θ and $\hat{\theta} = \hat{\theta}(D)$. In survival settings, the survival time Y_i may be censored at the time of analysis or when prediction is carried out; that is, we only know that Y_i exceeds the censoring time C_i. A frequentist framework is used to discuss the performance of predictive distributions. But Bayesian predictive distributions can be assessed under this framework as well.

A good probabilistic predictor $\hat{F}_p(y|z)$ should have two features, calibration and sharpness (Gneiting et al. 2007). Calibration means that the probability is "honest", in the sense that the fraction of $Y \leq y$ given $Z = z$ in future observations is close to $\hat{F}_p(y|z)$. Sharpness measures the degree of concentration of the predictive distribution $\hat{F}_p(y|z)$, which is equivalent to the sharpness of the corresponding predictive probability density function $\hat{f}_p(y|z) = d\hat{F}_p(y|z)/dy$. A narrow prediction interval can be expected when the plot of density $\hat{f}_p(y|z)$ vs. y is sharp-looking. This concept has a connection with prediction power of a point predictor, which will be shown later.

The notation used in the previous chapters is adopted here. First, we assume that there is a conceptual sequence of data sets $D_n = \{(y_i, z_i), i = 1 \ldots, n\}$ based on independent units (Y_i, Z_i). We assume further that given data D there is a procedure involving estimation and other aspects of model selection for obtaining a predictive distribution $\hat{F}_p(y|z)$. In addition, we assume that as $n \to \infty$, $\hat{F}_p$ converges at least pointwise to a limit F^*, for all z. Finally, the true distribution from which Y is generated given $Z = z$ is denoted $F_T(y|z)$, and the Y-values to be predicted correspond to Z-values arising from a distribution $H_Z(z)$. The Z-values in D may come from a different distribution or in some cases may be fixed by investigators.

6.1 Calibration and Sharpness

6.1.1 Calibration

Various forms of calibration have been proposed; these depend on whether predictive probabilities or prediction intervals are considered and whether D and Z are fixed or random. Let U(0, 1) denote the uniform distribution on the interval [0,1]. Then for continuous random variable Y, we call $\hat{F}_p(y|z)$ strongly calibrated if

$$U = \hat{F}_p(Y|z) \sim \mathrm{U}(0,\ 1), \tag{6.1}$$

for all $Z = z$ and where Y and D are both random. This differs from the strong calibration definition of Gneiting et al. (2007), who do not consider explicitly covariates Z or specification of $\hat{F}_p$. It is generally impossible to verify (6.1) when Z has continuous components, and a weaker form of calibration is then

$$U = \hat{F}_p(Y|Z) \sim \mathrm{U}(0,\ 1), \tag{6.2}$$

where $Y,\ Z$ and D are all random.

Standard model checking procedures involve (6.1) and (6.2). The uniform probability plots of residuals $\hat{u}_i = F_{\hat{\theta}}(y_i|z_i)$ provide checks of (6.2) on fitted models $F_{\hat{\theta}}$. We can also plot $\hat{u}_i$ against z_i or generate probability plots within strata defined by Z_i. When probabilistic prediction based on $\hat{F}_p$ is the objective, one may prefer deletion residuals $\hat{u}_i = \hat{F}_{p(-i)}(y_i|z_i)$, where $\hat{F}_{p(-i)}$ is based on the data $D/(y_i, z_i)$ with (y_i, z_i) dropped, similar to leave-one-out cross-validation. Other cross-validation residuals, or whenever possible, residuals $\hat{u}_j = \hat{F}_p(y_j|z_j)$ based on independent "test" data (Y_j, Z_j) could be valuable, too.

The calibration discussed above is termed probabilistic calibration by Gneiting et al. (2007), and we call (6.2) unconditional probabilistic calibration. In the same paper, Gneiting et al. discussed two other types of calibration. One of them is marginal calibration, which compares the observed and predicted Y-frequencies

$$\hat{F}_0(y) = \frac{1}{m}\sum_{j=1}^{m} \hat{F}_p(y|z_j) \ vs. \ \tilde{F}_0(y) = \frac{1}{m}\sum_{j=1}^{m} I(y_j \leq y), \tag{6.3}$$

where (y_j, z_j) are test data.

Many authors consider calibration under the assumption that the family $F_\theta(y|z)$ includes the true distribution $F_T(y|z)$; that is, $F_T(y|z) = F_{\theta_0}(y|z)$ for some θ_0. In that case one can seek a predictive distribution $\hat{F}_p(y|z)$ so that (6.1) is true, either exactly or approximately. If $\hat{\theta}_n = \hat{\theta}(D_n)$ is a consistent estimator of θ_0 as $n \to \infty$, then the standard "plug-in" choice $\hat{F}_p = F_{\hat{\theta}_n}$ yields (6.1) asymptotically. For finite n, a number of authors have discussed how to make (6.1) exact, or at least more accurate (e.g. Lawless and Fredette 2005 and references therein) in the case of classical settings where covariates are not considered.

6.1.2 Sharpness

Broadly, sharpness refers to the degree of dispersion of the predictor $\hat{F}_p(y|z)$ (Gneiting et al. 2007). It can be quantified or assessed by the standard deviation of the predictive distribution or the difference in its quantiles such as $\hat{Q}(0.95|z) - \hat{Q}(0.05|z)$, where $\hat{Q}(\alpha|z)$ satisfies $\hat{F}_p(\hat{Q}(\alpha|z)) = \alpha$. Note that $(\hat{Q}(\alpha/2|z), \hat{Q}(1-\alpha/2|z))$ gives a central $1-\alpha$ prediction interval for Y given z. When two or more well-calibrated predictive distributions are compared, the one with shorter prediction intervals is deemed to be better. The sharpness of a predictive distribution is naturally constrained by the sharpness of $F_T(y|z)$ and thus depends on the explanatory or prediction power of the covariates Z for Y. Numerous authors have noted that for survival models the predictive power of the covariates is often rather low (e.g. Korn and Simon 1990, Graf et al. 1999, Henderson et al. 2001, Schumacher et al. 2003, Bair and Tibshirani 2004, Rosthøj and Keiding 2004, and Henderson and Keiding 2005).

The performance of probabilistic predictors $\hat{F}_p(y|z)$, therefore, is determined by a combination of its calibration to $F_T(y|z)$, its sharpness, and the predictive power or sharpness of $F_T(y|z)$. Performance scores, also known as scoring rules, address these features simultaneously and have been used as summary measures for assessing the performance of predictive distributions.

6.2 Performance Scores

The two most common performance scores are the integrated Brier score (IBS) and logarithmic score (LS). The IBS is defined as

$$\mathrm{IBS}(Y) = \int_{-\infty}^{\infty} \{I(Y \leq s) - \hat{F}_p(s|z)\}^2 ds = \int_{-\infty}^{\infty} \mathrm{BS}(Y; s) ds, \tag{6.4}$$

where $\mathrm{BS}(Y; s) = \{I(Y \leq s) - \hat{F}_p(s|z)\}^2$ is called the Brier score at time s. $\mathrm{BS}(Y; s)$ defined here is actually one half the traditional Brier score given by Brier (1950) and is the squared error loss between the indicator variable $I(Y \leq s)$ and $E_{\hat{F}_p}[I(Y \leq s)]$.

The LS is defined as

$$\mathrm{LS}(Y) \quad = \quad -\log \hat{f}_p(Y|z), \tag{6.5}$$

where $\hat{f}_p(y|z) = d\hat{F}_p(y|z)/dy$ is the predictive density function.

Given m individuals with Z-values $z_1, \ldots, z_m$ and realized Y-values $y_1, \ldots, y_m$, the performance of a predictive distribution can be measured by the average IBS or LS,

$$\overline{\text{IBS}} = \frac{1}{m}\sum_{j=1}^{m} \text{IBS}(y_j), \tag{6.6}$$

$$\overline{\text{LS}} = \frac{1}{m}\sum_{j=1}^{m} \{-\log \hat{f}_p(y_j|z_j)\}. \tag{6.7}$$

We may also consider the expected IBS or LS, defined respectively as

$$\text{EIBS} = E\{\text{IBS}(Y)\},\ \text{ELS} = E\{\text{LS}(Y)\}. \tag{6.8}$$

The expectation is taken with respect to both Y and D, and often also with respect to a distribution $H_Z(z)$ for Z, because we are interested in assessing the performance of $\hat{F}_p$ obtained through a modeling procedure M. The definitions of EIBS and ELS are similar to the definition of prediction error $\pi_3(M; F_T, H_Z)$ in section 1.2. EIBS is called the continuous ranked probability score (CRPS) by Gneiting et al. (2007) and others. Similar to prediction error π, the smaller the EIBS or ELS, the better the predictor.

The scoring rules address both calibration and sharpness, which can be seen by decomposing the EIBS and ELS as follows:

$$\begin{aligned} \text{EIBS} &= E_Z \int_{-\infty}^{\infty} F_T(y|z)(1 - F_T(y|z))dy \\ &\quad + E_Z \int_{-\infty}^{\infty} E_D\{\hat{F}_p(y|z) - F_T(y|z)\}^2 dy \qquad (6.9) \\ \text{ELS} &= E_Z E\{-\log f_T(Y|z)\} + E_Z E[\log\{f_T(Y|z)/\hat{f}_p(Y|z)\}]. \qquad (6.10) \end{aligned}$$

The first terms in (6.9) and (6.10) are measures of sharpness of $F_T(y|z)$ which depend on the inherent variation of Y given Z. The second terms can be interpreted as bias or goodness of fit measures, which capture the discrepancy between $\hat{F}_p$ and F_T. In fact, the second term in (6.10) is the well-known Kullback-Leibler distance, which gives the distance from f_T (density of F_T) to the predictive density $\hat{f}_p$ averaged over the Z distribution. Both (6.9) and (6.10) are minimized when $\hat{F}_p = F_T$ for all Z, with the lower limits given by the first terms.

The first term of EIBS is less transparent and we give some connection with other measures. First we prove the following result, which links the first term in (6.9) with the length of prediction intervals.

Theorem 6.2.1 *Let $Q(p|z) = F_T^{-1}(p|z)$ denote the p-quantile of Y given $Z = z$ $(0 < p < 1)$, so that $L(\alpha; z) = Q(0.5 + 0.5\alpha \mid z) - Q(0.5 - 0.5\alpha \mid z)$ is the length of the central α prediction interval for Y given z $(0 < \alpha < 1)$. Assuming $E(Y|Z)$ exists, we have*

$$\int_{-\infty}^{\infty} F_T(y|z)[1 - F_T(y|z)]dy = \int_0^{0.25} L(\sqrt{1-4v}; z)dv. \tag{6.11}$$

Proof: For notational convenience we will suppress z in the following development. Let $y_L(v)$ and $y_U(v)$ be the smaller and larger of the two y-values satisfying $F(y)[1 - F(y)] = v$. When $F(y)$ is a continuous distribution function, we can verify that

$$F(y_L(v)) = 0.5 - 0.5\sqrt{1-4v} \text{ and } F(y_U(v)) = 0.5 + 0.5\sqrt{1-4v}.$$

for $0 < v \leq 0.25$. By the definition of $Q(p)$ and $L(\alpha)$, we can show

$$y_U(v) - y_L(v) = Q(0.5 + 0.5\sqrt{1-4v}) - Q(0.5 - 0.5\sqrt{1-4v}) = L(\sqrt{1-4v}),$$

which is illustrated in Figure 6.1. Therefore, provided that $E(Y)$ exists

$$\int_{-\infty}^{\infty} F(y)[1 - F(y)]dy = \int_0^{0.25} L(\sqrt{1-4v})dv.$$

Note that the integral on either side of equation (6.11) represents the area under the curve as shown in Figure 6.1. The integral on the left hand side corresponds to integrating from left to right with respect to y and the integral on the right hand side corresponds to integrating with respect to v from 0 to 0.25. □

Gneiting et al. (2007) suggest using both a Brier score plot and the boxplot of widths of prediction intervals for the visual comparison of the probabilistic predictors. The above theorem shows that the two types of plot are equivalent.

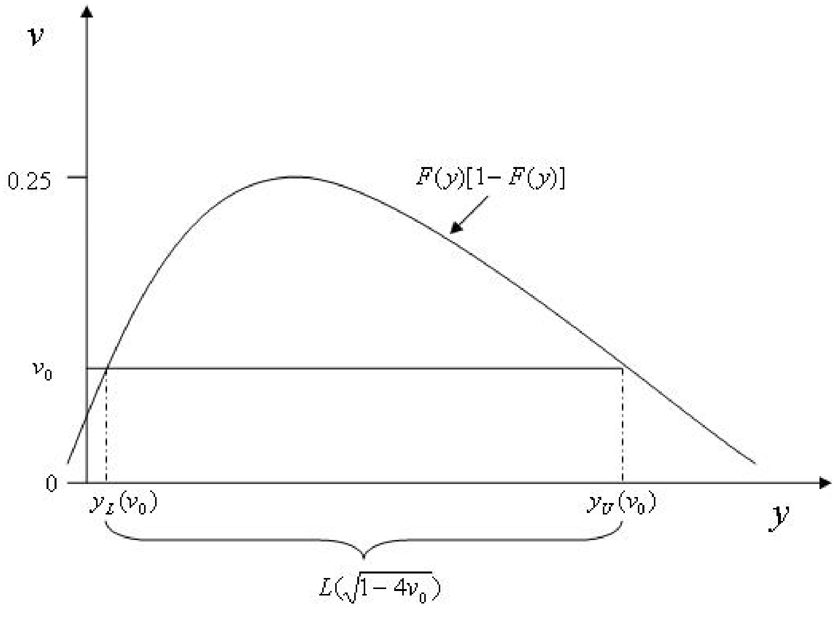

Figure 6.1: Illustration of Theorem 6.2.1.

Links can also be established between prediction error π for point predictors and expected performance scores for predictive distributions, due to the fact that π are in essence measures of variation around the point predictor. A connection for the first term of EIBS with expected absolute error loss is given by the following theorem.

Theorem 6.2.2 *Let $m(Z)$ denote the median of Y given Z under $F_T(y|z)$. Then*

$$0.5E\{|Y - m(z)|\} \leq \int_{-\infty}^{\infty} F_T(y|z)[1 - F_T(y|z)]dy \leq E\{|Y - m(z)|\}. \tag{6.12}$$

Again, for notational convenience, we suppress z in the following proof.

Proof: It is easily seen that for any distribution function $F(y)$,

$$0.5 \min(F(y), 1 - F(y)) \leq F(y)[1 - F(y)] \leq \min(F(y), 1 - F(y)),$$

in addition, integration by parts gives

$$\begin{aligned} \int_{-\infty}^{\infty} \min(F(y), 1 - F(y))dy &= \int_{-\infty}^{m} F(y)dy + \int_{m}^{\infty} (1 - F(y))dy \\ &= \int_{-\infty}^{m} (m - y)f(y)dy + \int_{m}^{\infty} (y - m)f(y)dy \\ &= E\{|Y - m|\} \end{aligned}$$

□

This provides a link between the expected absolute error loss and (6.11). It can also be shown along similar lines that

$$\int_{-\infty}^{\infty} F_T(y|z)[1 - F_T(y|z)]dy = E\{|Y - m(z)||1 - 2F(Y|z)|\},$$

which gives (6.11) explicitly as an expected loss corresponding to the loss function $L(Y, \hat{Y}) = |Y - \hat{Y}||1 - 2F(Y|z)|$ where $\hat{Y} = m(z)$.

6.3 Estimation of Expected Performance Scores

When "test" data $\{(y_j, z_j), j = 1, \dots, m\}$ is available for the assessment of $\hat{F}_p(y|z)$, the average integrated Brier score or logarithmic score over the predictions can be computed as in (6.6) and (6.7). However, when the test data is not available, it is often of interest to estimate the expected performance scores, i.e. EIBS or ELS (6.8). Graf et al. (1999) and Gerds and Schumacher (2006 & 2007) have recently considered this for the EIBS and also for the expected Brier score, $\text{EBS}(s) = E[\text{BS}(Y; s)]$, in the case of survival time prediction. They provide point estimates, but these are typically subject to substantial sampling variation, as are the prediction error estimates. We use the bootstrap method for confidence interval estimation (see section 6.4 below), similar to the confidence interval procedures for prediction error given in section 2.4. The procedures can be applied to either the EIBS or ELS.

In survival settings, the followup time is usually limited. Therefore, we consider truncated versions of IBS(Y), LS(Y) and their expectations. In particular, let τ denote an upper limit or followup time. We consider

$$\text{IBS}(Y; \tau) = \int_0^\tau \{I(Y \le s) - \hat{F}_p(s|z)\}^2 ds = \int_0^\tau \text{BS}(Y; s) ds \tag{6.13}$$

and

$$\text{LS}(Y; \tau) = I(Y \le \tau)\{-\log \hat{f}_p(Y|z)\} + I(Y > \tau)\{-\log \hat{S}_p(\tau|z)\}, \tag{6.14}$$

where $\hat{S}_p(\tau|z) = 1 - \hat{F}_p(\tau|z)$ is the predictive survivor function. The expectations of (6.13) and (6.14) are then

$$\text{EIBS}(\tau) = E\{\text{IBS}(Y; \tau)\}, \text{ and } \text{ELS}(\tau) = E\{\text{LS}(Y; \tau)\}. \tag{6.15}$$

Here the expectations are with respect to Y, Z and the data D giving $\hat{F}_p(y|z)$. Graf et al. (1999) and Gerds and Schumacher (2006 & 2007) considered bootstrap, cross-validation or simple plug-in estimates of the EIBS(τ) based on data D. An important contribution they made is the introduction of IPCW to address the censored BS($y; s$), as discussed in section 2.2.1 and this is briefly reviewed below.

The notation used here are similar to those in section 2.2.1, let Y denote the survival time and C denote the censoring time. The observed data D on n independent individuals consist of $\{(T_i, \delta_i, Z_i),\ i = 1, \ldots, n\}$, where $T_i = \min(Y_i, C_i)$, and $\delta_i = I(Y_i \leq C_i)$. Let

$$\Delta_i(s) = I(\text{BS}(Y_i; s) \text{ is observed}).$$

Note that $\Delta_i(s) = 1$ if $\delta_i = 1$ or if $\delta_i = 0$ but $C_i > s$. To apply IPCW, we assume independent censoring, and that

$$\hat{S}^c(s|z_i) = \widehat{Pr}(C_i > s \mid z_i),\ i = 1, \ldots, n,$$

provides consistent estimators of the $Pr(C_i > s \mid z_i)$. A naive or "plug-in" IPCW estimate of EIBS(τ) for the case in which Z has a uniform distribution on the values $\{z_1, \ldots, z_n\}$ observed in D is then (Gerds and Schumacher, 2006 & 2007)

$$\begin{aligned} \widehat{\text{EIBS}}(\tau) &= \int_0^\tau \widehat{\text{EBS}}(s)ds \\ &= \int_0^\tau \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i(s)}{\alpha_i(s)}\text{BS}(y_i; s)ds, \end{aligned} \tag{6.16}$$

where $\alpha_i(s) = S^c(y_i \wedge s|z_i) = Pr(\Delta_i(s) = 1 \mid y_i, z_i)$. Assuming that $S^c(\tau|z) > 0$ for all z, (6.16) provides a consistent estimator of EIBS(τ) as $n \to \infty$ (Gerds and Schumacher 2006).

The same IPCW approach can be applied to the estimation of ELS(τ) when some LS(y_i, τ) are censored. Define

$$\Delta_{1i} = I(Y_i \leq \tau, Y_i \leq C_i) \text{ and } \Delta_{2i} = I(Y_i > \tau, C_i > \tau),$$

then $\alpha_i = S^c(Y_i \wedge \tau|z_i)$ and the estimator of ELS(τ) is

$$\widehat{\text{ELS}}(\tau) = \frac{1}{n}\sum_{i=1}^n \left\{\frac{\Delta_{1i}}{\alpha_i}[-\log \hat{f}_p(y_i|z_i)] + \frac{\Delta_{2i}}{\alpha_i}[-\log \hat{S}_p(\tau|z_i)]\right\}. \tag{6.17}$$

The plug-in estimator (6.16) and (6.17) may be expected to underestimate the true expected performance score with respect to new data, because they are based on the training data D that was used to obtain $\hat{F}_p(y|z)$. This is similar to the underestimation of apparent loss for prediction error discussed in section 1.2.1. A cross-validation estimator (e.g. Gerds and Schumacher 2007) is a preferable alternative, especially when n is not very large relative to model size. The V-fold cross-validation estimator is, for EIBS(τ), given by

$$\begin{aligned}\widehat{\mathrm{BS}}_{-v}(Y;s) &= \{I(Y \le s) - \hat{F}_{p(-v)}(s|z)\}^2, &(6.18)\\ \widehat{\mathrm{EIBS}}^{cv}(\tau) &= \int_0^\tau \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v}\frac{\Delta_i(s)}{\hat{\alpha}_i(s)}\mathrm{BS}_{-v}(y_i;s)ds &(6.19)\end{aligned}$$

where $\hat{F}_{p(-v)}(y|z)$ is based on D with S_v left out. Note that $\hat{\alpha}_i(s)$ is still estimated from the complete data D, as suggested by Gerds and Schumacher (2007). The cross-validation version of the ELS(τ) estimator could be similarly defined as

$$\widehat{\mathrm{ELS}}^{cv}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\Delta_{1i}}{\hat{\alpha}_i}[-\log \hat{f}_{p(-v)}(y_i|z_i)] + \frac{\Delta_{2i}}{\hat{\alpha}_i}[-\log \hat{S}_{p(-v)}(\tau|z_i)]\right\}. \quad (6.20)$$

6.4 Construction of Confidence Intervals

In section 2.4, we used bootstrap methods for the construction of confidence intervals for prediction error, which worked well as demonstrated by the simulation studies in chapters 3 and 4. We take a similar approach here for EIBS(τ) and the confidence interval procedures are based on the generation of B pseudo training samples D_b^* ($b = 1, \ldots, B$) by nonparametric bootstrap sampling. Parametric bootstrap for generating D_b^* is not considered due to the sensitivity of the parametric bootstrap to model misspecification, as discussed in chapter 3.

The confidence interval procedures based on the V-fold cross-validation estimator $\widehat{\mathrm{EIBS}}^{cv}(\tau)$, abbreviated $\widehat{\mathrm{EIBS}}$, is discussed here for illustration. We need B bootstrap estimates $\widehat{\mathrm{EIBS}}_b^*$, each from a pseudo training data set D_b^*, for the estimation of var($\widehat{\mathrm{EIBS}}$). The procedure starts by drawing n items from the original training data $D = \{(y_i, z_i), i = 1, \ldots, n\}$ with replacement, to give D_b^*. Then we divide each D_b^* into S_v^*, V sets of approximately equal sizes, and apply the modelling procedure on D_b^* with S_v^* omitted. The resulting predictive distributions $\hat{F}^*_{p(-v)}(y|z)$

are used to give estimates of $\widehat{\mathrm{BS}}^*_{-v}(y_i; s)$ for $y_i \in S^*_v$ as in (6.18) and then $\widehat{\mathrm{EIBS}}^*_b$ as in (6.19). Note that $\hat{\alpha}^*_i(s)$ in (6.19) is estimated from the complete data D^*_b.

The variance of $\widehat{\mathrm{EIBS}}$ is estimated with the bootstrap estimates $\widehat{\mathrm{EIBS}}^*_b$, $b = 1, \ldots, B$,

$$\widehat{\mathrm{var}}(\widehat{\mathrm{EIBS}}) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\mathrm{EIBS}}^*_b - \overline{\widehat{\mathrm{EIBS}}}^*)^2,$$

where $\overline{\widehat{\mathrm{EIBS}}}^* = \sum_{b=1}^{B}\widehat{\mathrm{EIBS}}^*_b/B$. We then treat $(\widehat{\mathrm{EIBS}} - \mathrm{EIBS})/\sqrt{\widehat{\mathrm{var}}(\widehat{\mathrm{EIBS}}^*)}$ as a normally distributed pivotal quantity, producing $1-\alpha$ confidence intervals $\widehat{\mathrm{EIBS}} \pm \Phi^{-1}(\alpha/2)\sqrt{\widehat{\mathrm{var}}(\widehat{\mathrm{EIBS}}^*)}$, where $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal distribution and $\Phi^{-1}(\alpha/2)$ gives the $\alpha/2$ quantile for the standard normal distribution.

6.5 An Example

For illustration, we use the Mayo PBC data and consider the Weibull, Lognormal and Cox PH models with variable selection (see section 5.2 for details). The performance of the probabilistic predictors given by the fitted models is assessed with the methods discussed above.

We first look at the calibration. Figure 6.2 shows unconditional probabilistic calibration and marginal calibration plots of the three models. The probabilistic calibration plot is a uniform probability plot of residuals. Let $\hat{u}_{(i)}$, $i = 1, \ldots, n$ denote the ordered $\hat{u}_i = \hat{F}_p(y_i|z_i)$; the uniform probability plot plots $E(U_{(i)}) = i/(n+1)$ under the U(0, 1) hypothesis vs. $\hat{u}_{(i)}$. Points that follow a straight line with intercept 0 and unit slope suggest a well calibrated probabilistic predictor $\hat{F}_p(y_i|z_i)$. For a censored survival time y_i, its corresponding $\hat{u}_i = \hat{F}_p(y_i|z_i)$ is also censored and the Kaplan-Meier estimate of $Pr(U > \hat{u}_i)$ is used. We then plot the estimated $\widehat{Pr}(U \leq \hat{u}_i)$ vs. the observed $\hat{u}_i$. Note that the $\hat{u}_i$ plotted in the left panels of Figure 6.2 are the leave-one-out cross-validation (LOOCV) estimates $\hat{F}_{p(-i)}(y_i|z_i)$. The graphs indicate that the three models are satisfactory in terms of unconditional probabilistic calibration (6.2),

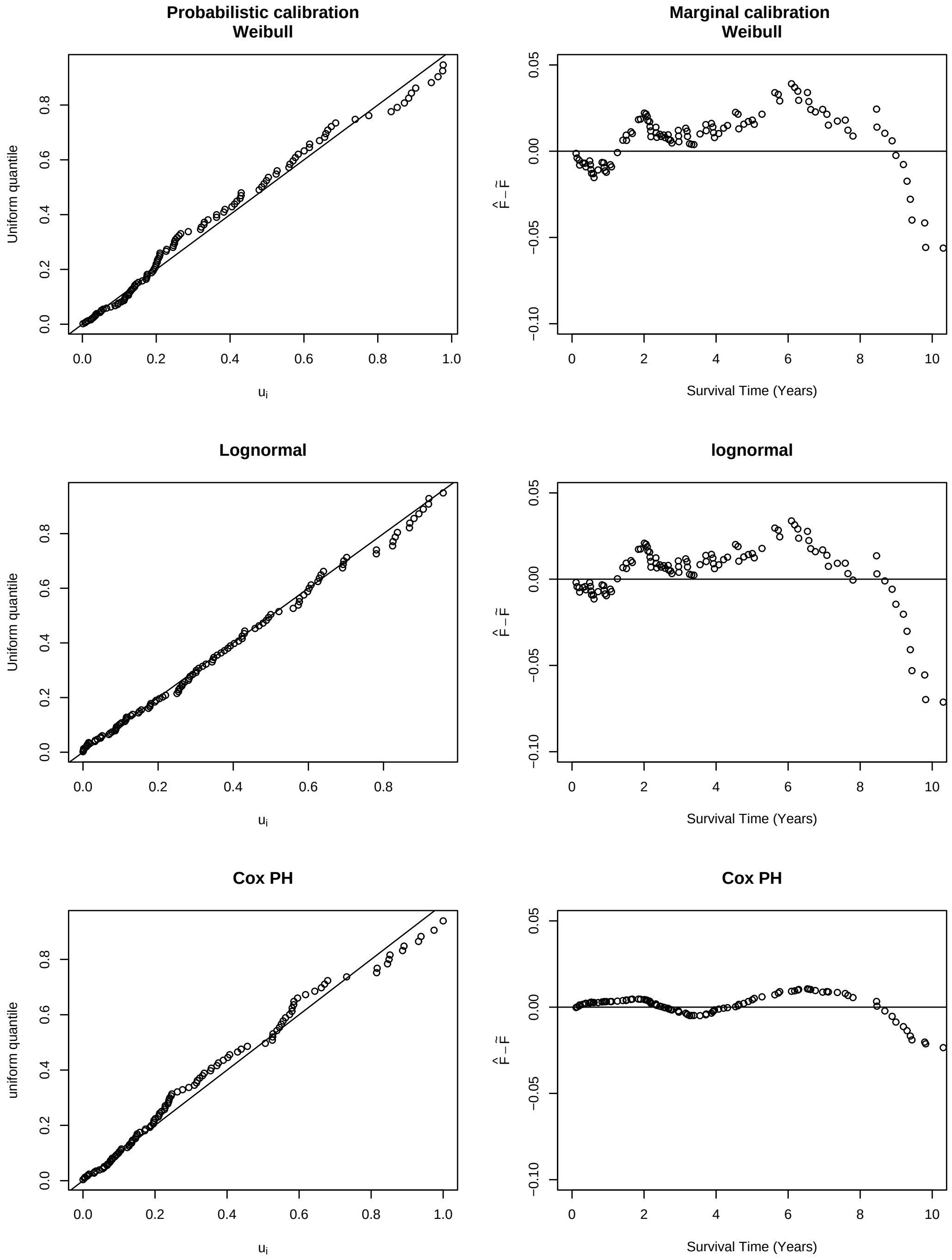

Figure 6.2: Calibration plots for the Weibull, Lognormal and Cox PH models on PBC data. The left panels are the unconditional probabilistic calibration plots and the right panels are the marginal calibration plots.

since there is no obvious departure from the reference lines. However, we should not read too much into the plot for the Cox PH model, because Cox PH is a semi-parametric model with a nonparametric component. When the effects of covariates are small, the u_i of Cox PH model are close to U(0,1) by construction (Lawless 2003, page 359).

A marginal calibration plot proposed by Gneiting et al. (2007) is adopted in Figure 6.2. It plots the difference of $\hat{F}_0(y)$ and $\tilde{F}_0(y)$, both defined in (6.3), as a function of time y. Note that our $\hat{F}_0(y)$ is a LOOCV estimate, given by $n^{-1}\sum_{i=1}^{n}\hat{F}_{p(-i)}(y_i|z_i)$ and $\tilde{F}_0(y)$ is the Kaplan-Meier estimate. Three such plots are made, for the Weibull, Lognormal and Cox models, respectively. They are displayed in the right panels of Figure 6.2. Under the hypothesis of marginal calibration, we expect to see minor fluctuations around the horizontal line at 0. The plots show small oscillations around 0 for small Y, but as Y increases, the variability in $\hat{F}_0(y) - \tilde{F}_0(y)$ also increases. This pattern holds true for all three models. It may be partially due to the fact that the variance of the estimated marginal probability $\tilde{F}_0(y)$ increases as y increases. Intuitively, as time passes and more subjects are censored, we have fewer subjects in the sample and the estimate of marginal survival probability becomes less precise. We also notice that the marginal calibration plots for Lognormal and Weibull models have very similar patterns, indicating that the $\hat{F}_0(y)$ estimates are very close between these two models. Cox PH model shows the best marginal calibration with the empirical $\tilde{F}_0(y)$ among the three competing models. This is somewhat due to the semiparametric nature of Cox PH model. Note that if we fit a nonparametric model to the marginal distribution of Y, e.g. a Kaplan-Meier estimate for $\hat{F}_p(y)$, then we have perfect marginal calibration as well as unconditional probabilistic calibration.

The point estimates of EIBS(τ) (for $\tau = 10$ years) and the confidence intervals for the three competing models are summarized in Table 6.1. Nonparametric bootstrap with $B = 500$ is used for confidence interval estimation. The LOOCV estimates of EIBS(τ) are greater than their plug-in counterparts, both sets of estimates are contained in the confidence intervals of each other. The LOOCV estimates of EIBS(τ) of Weibull and Cox PH model are close. The

	Weibull		Lognormal		Cox PH	
	$\hat{\pi}$	95% CI	$\hat{\pi}$	95% CI	$\hat{\pi}$	95% CI
Plug-in	1.06	(0.88, 1.23)	1.12	(0.93, 1.35)	1.01	(0.82, 1.20)
LOOCV	1.14	(0.94, 1.30)	1.22	(1.00, 1.43)	1.12	(0.90, 1.34)

Table 6.1: The estimated expected integrated Brier score and 95% confidence interval for the PBC data under the Weibull, Lognormal and Cox PH models, truncated at $\tau = 10$ years. The confidence intervals are based on 500 nonparametric bootstrap samples and the approximation that $\widehat{\text{EIBS}}$ are normally distributed.

confidence intervals under Cox PH model are slightly wider than under Weibull model. The estimated EIBS(τ) of Lognormal model (both plug-in and LOOCV estimates) are the greatest, but they are still located inside the respective confidence intervals of the other two models. Overall, these results suggest that the predictive performance of the three competing models are similar for the PBC data. The EIBS(τ) of the marginal distribution of Y, which corresponds to the Kaplan-Meier estimate $\tilde{F}_0(y)$, often serves as a reference for model-based probabilistic predictors (e.g. Schumacher et al. 2003). For the PBC data, the estimated marginal EIBS(τ) is 1.72 and the corresponding confidence interval is (1.55, 1.89), much greater than the estimates of the regression models reported in Table 6.1.

Figure 6.3 plots the estimated expected Brier score $\widehat{\text{EBS}}(y)$ for the three associated models. The two graphs in the top panels and the one in the bottom left panel illustrate the integral representation on the right hand side of the plug-in estimator (6.16) and the LOOCV estimator (6.19) of $\widehat{\text{EBS}}(y)$, respectively for Weibull, Lognormal and Cox PH models. The last graph in the bottom right panel contrasts the $\widehat{\text{EBS}}(y)$ of the marginal model with the LOOCV estimates of the three regression models. Note that for Cox PH model and the marginal model, the expected Brier scores are evaluated at distinct failure times and $\widehat{\text{EBS}}(y)$ are step functions of y. On the other hand, the expected Brier scores for Weibull and Lognormal models are estimated at 100 evenly spaced time points between 0 and 10 years; the graphs consist of line segments that connect neighboring points.

As expected, we find that the cross-validation estimates are slightly greater than the plug-in estimates for all three regression models (top panels and bottom left panel in Figure 6.3). The last graph indicates that $\widehat{\text{EBS}}(y)$ of Lognormal model is a bit larger than those of Cox PH and Weibull models, which are close. Since a smaller score indicates better performance, we may conclude that the Cox PH and Weibull model outperform the Lognormal model on the PBC data in terms of probabilistic prediction. However, the difference in performance is small and very likely insignificant, as suggested by the largely overlapping confidence intervals of $\text{EIBS}(\tau)$ of the three models (Table 6.1). The line representing $\widehat{\text{EBS}}(y)$ of the marginal model, which lies above the other lines of the regression models, is plotted in the same graph as a reference. For some region of y, e.g. year 3 to 6, the difference of $\widehat{\text{EBS}}(y)$ between the marginal model and the regression models is large, which suggests superior predictive performance of the probabilistic predictors given by the regression models. We further note that this graph is similar to Figure 5.6, which shows the misclassification error of the marginal model and Weibull model at selected time points.

The expected logarithm score $\text{ELS}(\tau)$ and the associated confidence intervals could also be estimated, using (6.17) and the confidence interval procedures described above, for the PBC example and the three models. But whether $\text{ELS}(\tau)$ of different models with possibly different model sizes are truly comparable remains an open question. Simulation studies to examine the width and coverage of prediction intervals, the estimators of $\text{EIBS}(\tau)$ and $\text{ELS}(\tau)$, and the performance of the proposed confidence interval procedures would be worthwhile. This investigation will be carried out in the future.

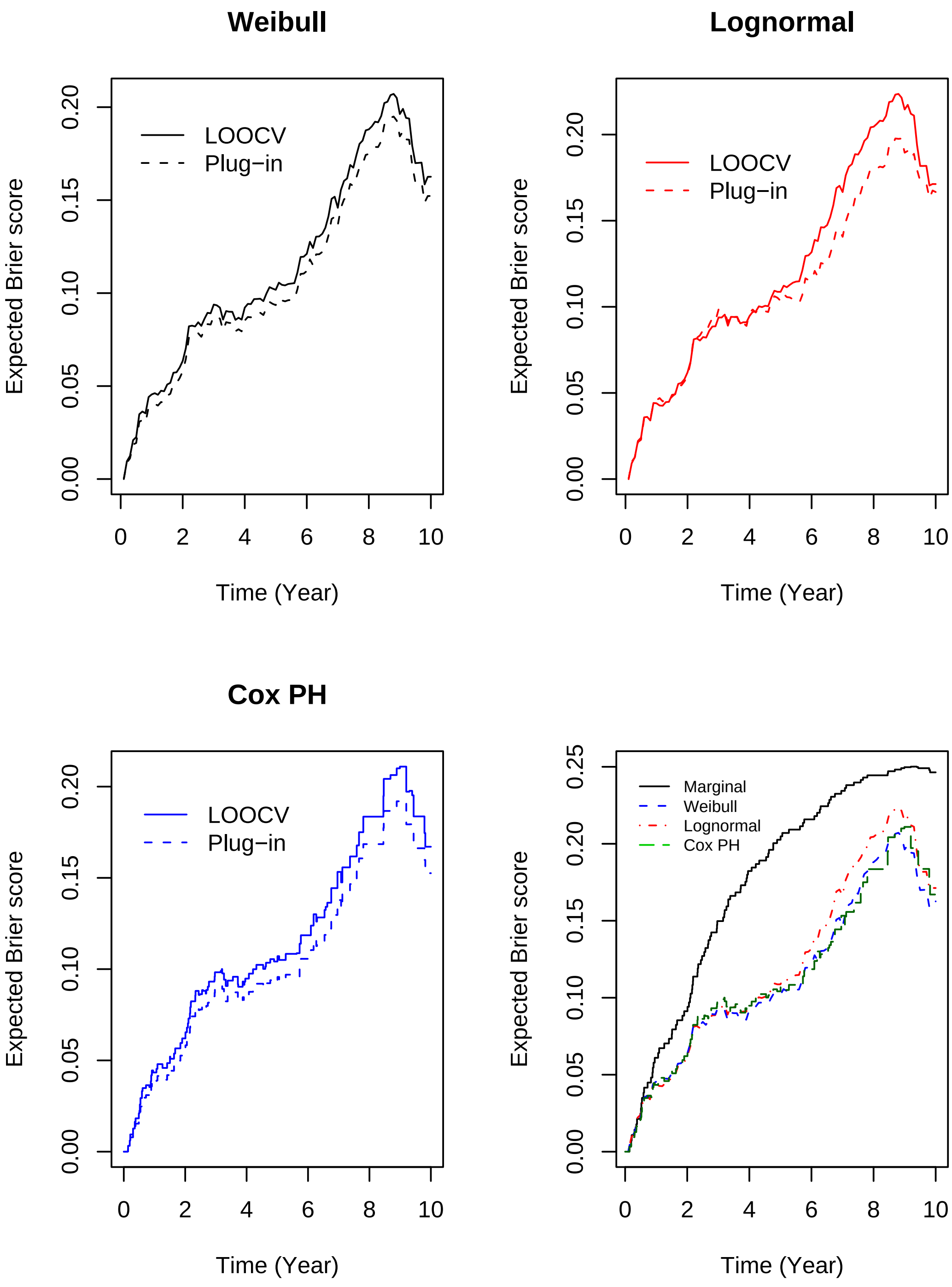

Figure 6.3: Expected Brier score plot on PBC data. The plug-in (dashed line) and leave-one-out cross-validation (LOOCV, solid line) estimates are plotted in the top panels and the bottom left panel for Weibull, Lognormal and Cox PH model, respectively. The graph in the bottom right panel shows the estimated expected Brier score for the marginal model with the LOOCV estimates of the three regression models.

Chapter 7

Future Research Topics

7.1 Change of Covariates' Distribution

In previous chapters, we assume that the prediction is made for Y arising from the same joint distribution (Y, Z) where the training data D is from. That is, both the conditional distribution function $F_T(y|z)$ and the marginal distribution function $H_Z(z)$ stay stationary. Moreover, when the estimation of prediction performance is considered, we further assume that Z is uniformly distributed on the values $(z_1, \ldots, z_n)$ observed in D, see (1.4), (6.16) and (6.17). This very restrictive assumption is frequently made by other authors, either explicitly (e.g. Efron 2004) or implicitly (e.g. Gerds and Schumacher 2007). In this section, we relax the assumption on the distribution of Z and consider the case where prediction of Y is made for $Z \sim H_M(z)$, different from the Z distribution $H_D(z)$ in the training data. We call it the changing covariates' distribution problem, which naturally applies to the prediction for future population where Z distribution is likely to evolve over time, e.g. the age structure in the population of patients with certain diseases.

Shimodaira (2000) studied this issue for probabilistic predictor. He considered a simple case where Z is univariate. Let $h_M(z) = dH_M(z)/dz$ denote the density function of the new Z for which prediction is to be made and $h_D(z) = dH_D(z)/dz$ denote the density function of covariates Z observed in D. By using the importance sampling identity, Shimodaira (2000) showed that when the weighting function $\xi(z) = h_M(z)/h_D(z)$ is applied to the observed sample for the estimation of θ, the resulting predictive distribution $F_{\hat{\theta}}(y|z)$ is optimal, in the sense that the expected logarithm score averaged over $H_M(z)$ is minimized.

We consider the $H_M \neq H_D$ situation for point predictor. Our objective is to assess the performance of predictor $\hat{Y}_D(Z) = G(Z; \hat{\theta})$, obtained by applying modeling procedure to training data D of size n that comes from $F_T(y|z)H_D(z)$. The prediction error of $\hat{Y}_D(Z)$ for $D_M = \{(y_j, z_j),\ j = 1, \ldots, m\}$ arising from $F_T(y|z)H_M(z)$ is of interest. That is, we want to estimate

$$\pi(M; F_T, H_M, H_D) = E_{H_M, F_T, H_D}[L(Y, \hat{Y}_D(Z))]. \tag{7.1}$$

For simplicity $\pi(M; F_T, H_M, H_D)$ is abbreviated as $\pi(H_M)$. Suppose such D_M is available, then the average prediction loss

$$\overline{PL} = \frac{1}{m} \sum_{j=1}^{m} L(y_j, \hat{y}_D(z_j))$$

provides a consistent estimator of $\pi(H_M)$. When covariates Z from $H_M(z)$ are available but Y is yet to be observed, $\pi(H_M)$ has to be estimated with training data D and $H_M(z)$.

To get an idea of how prediction error is affected by the change of covariates' distribution, consider a normal linear model with p covariates,

$$Y = Z\beta + \varepsilon, \quad \varepsilon \sim N(0,\ \sigma^2).$$

Let Z_D and Z_M denote the $n \times p$ and $m \times p$ covariate matrices associated with Y_D ($n \times 1$ vector) in the training data and Y_M ($m \times 1$ vector) to be predicted, respectively, and $\mu_M = E(Y_M|Z_M)$. We further assume that Z_D and Z_M are fixed. Using squared error loss, the optimal predictor is $\hat{Y}_D(Z_M) = Z_M\hat{\beta}_D$. It follows that

$$\begin{aligned}
\pi(H_M) &= \frac{1}{m}E_{F_T}[(Y_M - \hat{Y}_D(Z_M))^T(Y_M - \hat{Y}_D(Z_M))] \\
&= \frac{1}{m}E_{F_T}[(Y_M - \mu_M)^T(Y_M - \mu_M)] \\
&\quad + \frac{1}{m}E_{F_T}[(\hat{Y}_D(Z_M) - \mu_M)^T(\hat{Y}_D(Z_M) - \mu_M)] \\
&= \sigma^2 + \frac{1}{m}\text{trace}[\text{var}(\hat{\beta}_D)(Z_M^T Z_M)] \\
&= \sigma^2 + \frac{\sigma^2}{m}\text{trace}[(Z_D^T Z_D)^{-1}(Z_M^T Z_M)]. \qquad (7.2)
\end{aligned}$$

When $Z_M = Z_D$, (7.2) simplifies and $\pi(H_M) = (1 + p/m)\sigma^2$; when $Z_M \neq Z_D$, $\pi(H_M)$ depends on the relationship between Z_D and Z_M, and has to be evaluated numerically. This example indicates that when estimating prediction error, the covariates' distribution should be accounted for, even when the error distribution of Y given Z is the same for all z.

In section 1.2.1, we discussed three approaches, namely the model-based, the apparent loss based and the cross-validation methods, for the estimation of $\pi(M; F_T, H_D)$. When $H_M \neq H_D$, we note that the model-based method can be applied directly with minor adjustments. The apparent loss and cross-validation based methods require further consideration and we offer some preliminary thoughts and proposals.

7.1.1 Model-based method

Assume sets $(z'_1, \ldots, z'_m)$ and $(z_1, \ldots, z_n)$, respectively coming from H_M and H_D, are given. The empirical distributions based on the two sets are denoted by $\tilde{H}_M$ and $\tilde{H}_D$ henceforth. The model-based method uses an estimator $\hat{F}_T$, from which D given $\tilde{H}_D$ and D_M given $\tilde{H}_M$ are assumed to arise. The estimator is then

$$\hat{\pi}^m(\tilde{H}_M) = \hat{\pi}(M; \tilde{H}_M, \hat{F}_T, \tilde{H}_D) = \frac{1}{m}\sum_{j=1}^{m} E_D \int_{A_Y} L(y, G(z'_j; \hat{\theta}))d\hat{F}_T(y|z'_j), \qquad (7.3)$$

where A_Y denotes the range of Y. A popular choice for $\hat{F}_T$ is the fitted model $F_{\hat{\theta}}$. We commented in section 1.2 that $\hat{\theta} = \hat{\theta}(D)$ is a complex function of training data D and therefore an analytical estimate of $\pi(\tilde{H}_M)$ is usually not feasible. We can use a simulation procedure as shown below to evaluate (7.3):

1. Apply the prediction procedure M to D, giving an estimated model $F_{\hat{\theta}}(y|z)$.

2. Generate pseudo data $D^{*k} = \{(y_1^*, z_1), \ldots, (y_n^*, z_n)\}$ and $D_M^{*k} = \{(y_1^{'*}, z_1'), \ldots, (y_m^{'*}, z_m')\}$ from $F_{\hat{\theta}}(y|z)$.

3. Apply the modeling procedure to D^{*k}, which gives predictor $\hat{y}_{D^{*k}} = G(z; \hat{\theta}_k^*)$. With loss function $L^\tau(Y, \hat{Y})$, the average prediction loss for D_M^{*k} is

$$PL_k = \frac{1}{m} \sum_{j=1}^{m} L^\tau(y_j^{'*}, G(z; \hat{\theta}_k^*)).$$

4. Repeat steps 2 and 3 K times, obtaining a model-based estimate of prediction error $\pi(\tilde{H}_M)$

$$\hat{\pi}^m(\tilde{H}_M) = \frac{1}{K} \sum_{k=1}^{K} PL_k.$$

For censored survival data, we modify the above procedures by generating censoring times, as described in section 2.4. If instead of specific sets $(z_1', \ldots, z_m')$ and $(z_1, \ldots, z_n)$, H_M and H_D are known, we can generate Z_M^* and Z_D^* randomly from the corresponding distribution functions for D_M^{*k} and D^{*k} in step 2.

A critical problem with this approach is that when the model is misspecified, the model-based estimator (7.3) converges to $\pi(M; H_M, F_{\theta^*}, H_D)$, rather than $\pi(M; H_M, F_T, H_D)$, under suitable regularity conditions. This can be shown along similar lines given in section 2.3.

7.1.2 Weighting the apparent loss and cross-validation estimators

The estimators based on apparent loss (1.16) and cross-validation method (1.17) converge to $\pi(M; F_T, H_D)$ asymptotically, thus they may not be appropriate estimators on their own for $\pi(H_M)$. Inspired by Shimodaira (2000), we consider importance sampling and obtain

$$\begin{aligned} \pi(H_M) &= E_{H_M,F_T,H_D}[L(Y, \hat{Y}_D(Z))] \\ &= E_{H_D} \int_{A_M} \int_{A_Y} L(y, \hat{y}_D(z)) F_T(y|z) h_M(z) dydz \\ &= E_{H_D} \int_{A_D \cup A_M} \int_{A_Y} \frac{h_M(z)}{h_D(z)} L(y, \hat{y}_D(z)) F_T(y|z) h_D(z) dydz \\ &= E_{H_D,F_T} \left[\frac{h_M(z)}{h_D(z)} L(y, \hat{y}_D(Z)) \right], \end{aligned} \tag{7.4}$$

where A_M denotes the range of Z in H_M. For (7.4) to hold, we require $A_D \supseteq A_M$ where A_D denotes the range of Z in H_D, so that $h_M(z)/h_D(z)$ exists for all $z \in A_D$. The last term $E_{H_D,F_T} \left\{ \frac{h_M(z)}{h_D(z)} L(y, \hat{y}_D(Z)) \right\}$ in (7.4) can be estimated by, for example, a weighted apparent loss based estimator

$$\hat{\pi}(\tilde{H}_M) = \frac{1}{n} \sum_{i=1}^{n} \xi_i L(y_i, \hat{y}_i) \tag{7.5}$$

with

$$\xi_i = \hat{h}_M(z_i)/\hat{h}_D(z_i), \tag{7.6}$$

This seems to be a sensible approach. Theoretically we could estimate the density of Z using kernel or projection pursuit method, but it is well-known that for moderate number of covariates, an accurate estimate of the multivariate density h_M or h_D is difficult to obtain, especially when Z has both continuous and categorical components. This might be the reason that in Shimodaira (2000) a single explanatory variable Z with normal density is considered.

The weighting approach suggested in (7.5) may still be a viable idea. The most appropriate weight for $L(y_i, \hat{y}_i)$, is given by $h_M(z_i)/h_D(z_i)$, the relative probability of observing z_i in H_M to observing it in H_D. Motivated by the above, we consider constructing the weights with respect to the relative location of z_i in $\tilde{H}_D$ and $\tilde{H}_M$, using some distance measure. In particular, we choose the Mahalanobis distance, which is defined as

$$d_{ij} = \sqrt{(z_i - z'_j)\Sigma_M^{-1}(z_i - z'_j)} \tag{7.7}$$

for points z_i in $\tilde{H}_D$ and z'_j in $\tilde{H}_M$. It measures the distance of z_i to z'_j scaled by the covariance matrix Σ_M of $\tilde{H}_M$. We use this distance because we think it is important to consider the correlation structure among the covariates when distance between vectors are measured.

Let $K(\cdot)$ denote a kernel function, we define

$$\eta_{ij} \;=\; \frac{K(d_{ij})}{\sum_{i=1}^{n} K(d_{ij})}.$$

η_{ij} is a measure of influence from z_i to z'_j relative to other points in $\tilde{H}_D$, and increases as z_i approaches z'_j. Summing η_{ij} over j, we get

$$\eta_i^M \;=\; \sum_{j=1}^{m} \eta_{ij} = \sum_{j=1}^{m} \frac{K(d_{ij})}{\sum_{i=1}^{n} K(d_{ij})}.$$

η_i^M gives the overall contribution of z_i to $\tilde{H}_M$. It increases as z_i approaches the high density area of $\tilde{H}_M$. Thus η_i^M gives a density-like estimate of z_i in $\tilde{H}_M$. Note that η_{ij} is defined such that $\eta_j^M = \sum_{i=1}^{n} \eta_{ij} = 1$ for all z'_j. This constraint ensures that each observation (y'_j, z'_j) contributes equally to the estimator $\pi(\tilde{H}_M)$. Similarly, we get the density-like estimate of z_i in $\tilde{H}_D$ with

$$\begin{aligned} d_{ii'} &= \sqrt{(z_i - z'_i)\Sigma_D^{-1}(z_i - z'_i)} \\ \eta_{ii'} &= \frac{K(d_{ii'})}{\sum_{i=1, i\neq i'}^{n} K(d_{ii'})}, \\ \eta_i^D &= \sum_{i'=1, i'\neq i}^{n} \eta_{ii'} = \sum_{i'=1, i'\neq i}^{n} \frac{K(d_{ii'})}{\sum_{i=1, i\neq i'}^{n} K(d_{ii'})}, \end{aligned}$$

and note that $\eta_{i'}^D = \sum_{i=1, i\neq i'}^{n} \eta_{ii'} = 1$.

The weight ξ_i in (7.5) is then estimated by

$$\xi_i = \eta_i^M / \eta_i^D. \tag{7.8}$$

Simulation studies are needed to examine the proposed distance-based nonparametric method.

We showed that for normal linear models with fixed number of covariates, the prediction error on D_M has a closed form

$$\pi(H_M) = \sigma^2(1 + \frac{1}{m}\text{trace}[(Z_D^T Z_D)^{-1}(Z_M^T Z_M)].$$

If we ignore the change in the distribution of Z and use $\pi = \sigma^2(1 + p/n)$ as the estimator for $\pi(H_M)$, the bias is

$$\text{Bias} = \pi - \pi(H_M) = \sigma^2\{\frac{p}{n} - \frac{1}{m}\text{trace}[(Z_D^T Z_D)^{-1}(Z_M^T Z_M)]\}$$

The magnitude of the bias depends on trace$[(Z_D^T Z_D)^{-1}(Z_M^T Z_M)]$. We conducted a simulation experiment to examine the bias and the performance of the proposed weighted estimator. The simulation model is

$$\log(Y) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon,\ \varepsilon \sim \text{EV}(0,\ 1),$$

where

$$H_D \sim \text{BVN}\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 10 & 1 \\ 1 & 4 \end{pmatrix}\right) \qquad H_M \sim \text{BVN}\left(\begin{pmatrix} 2.5 \\ 1.6 \end{pmatrix}, \begin{pmatrix} 4 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$$

The sample size is $n = 200$ for training data D and $m = 100$ for new data D_M. We use the absolute relative error loss ($L(Y, \hat{Y}) = |\log Y - \log \hat{Y}_m|$) to evaluate prediction performance and the Gaussian kernel function for estimation of the weights ξ given by (7.8). The simulation results suggest that the weighted estimator (7.5) is indeed less biased. But the bias of $\pi - \pi(H_M)$ is small and amounts to less than 4% of average($\hat{\pi}$). In addition, the variance of weighted estimator $\hat{\pi}(H_M)$ tends to be greater than that of $\hat{\pi}$ because $w_i > 1$ for some i. Thus, the weighting method does not offer substantial improvement under the current simulation setting. It is difficult to reduce the variance inflation, since it is inherent in the weighting method. We may, however, find a setting in which the bias of $\pi - \pi(H_M)$ is relatively large and justify the use of the weighted estimator. Possible settings include:

- Under the above simulation setting, we assess the prediction error with $L(Y, \hat{Y}) = |Y - \hat{Y}|$ on the original scale of Y instead of using the loss function on the $\log(Y)$ scale. This is because $\log(Y)$ given Z has the same error distribution, fixed by the simulation model. Therefore, $E_{F_T}[L(\log(Y), \log(\hat{Y})) \mid Z]$ are identical for all z. But $E[L(Y, \hat{Y}) \mid Z]$ under F_T depends on Z through the linear predictor $\beta^T Z$.

- We can consider assessing the prediction error with the truncated loss function $L^{\tau}(Y, \hat{Y})$. The truncated loss also depends on Z through the linear predictor $\beta^T Z$.

- We can also consider the case where some points in H_M are far away from the points in H_D, e.g. reversing the role of H_M and H_D in the above simulation setting.

7.2 Other Topics

7.2.1 Issues in prediction error and prediction power estimation

In chapters 2, 3 and 4, we studied the performance of point predictors allowing for model misspecification and variable selection, and proposed prediction error estimators for censored survival data. The model-based point and confidence interval estimators are shown to be sensitive to model misspecification, by which we mean the error distribution is wrongly specified. This is different from the model misspecification studied by Rosthøj and Keiding (2004), where they examine the misspecification of the linear predictor in the normal and the logistic regression models. They found that the model-based prediction power estimates are close to the true prediction power. Presumably under our definition of model misspecification, the model-based estimators of marginal prediction error and prediction power should behave like the model-based estimators for regression models and be biased estimators. This can be investigated in the future.

We studied the parametric models with simulation studies (chapter 3 and 4) and compared the prediction performance of parametric models and semi-parametric Cox PH model in the data analysis examples (chapter 5). Presumably, properties of prediction error and confidence interval estimates of the Cox PH model are similar to those of the parametric models, but investigation of this could be valuable. Predictors given by nonparametric modeling procedures such as regression trees and neural networks could also be examined under the same framework, however, their confidence intervals require more computation.

7.2.2 Prediction with time-varying covariates

Current literature and the present study limited the attention to the baseline covariate values and evaluated their prediction performance for short and long-term survival time. Realistically, it may not be reasonable to expect that the baseline values have much power in determining the long-term survival probabilities. Often during the followup period, patients are seen regularly and measurements of certain physiological variables are repeatedly taken to monitor the disease progression. Some of these time-varying physiological variables could be predictors of failure occurring in the near future. Thus it is important to model them for the prediction of short-term survival experiences and disease progressions. For example, a patient might be predicted to survive the next 6 months if,

$$\widehat{Pr}(Y_i > t + s \mid Y_i > t, Z_i(t)) > 0.5,$$

where $s = 6$ months, and $Z(t)$ represents the values of time-varying covariates at time t. Alternatively, we could use $\widehat{Pr}(Y_i > t + s \mid Y_i > t, Z_i(t))$ as a probabilistic predictor.

Taylor et al. (2005) modeled the baseline covariates along with the time-varying prostate-specific antigen (PSA) for prostate cancer patients and were successful in predicting future PSA values and times to clinical recurrences. It is our interest to investigate the assessment of predictors' performance in this ”dynamic” setting.

7.2.3 Updating and sequential validation of prediction performance

In survival studies, patients are often recruited for a period of time. Their survival times could be highly variable and sometimes may be very large. Therefore, a long study or followup time may be needed to obtain performance measures for statistical models. These measures include prediction error (1.3) for point predictors, expected integrated Brier score (EIBS) (6.6) and expected logarithm score (ELS) (6.7) for probabilistic predictors, or the truncated versions like EIBS(τ) and ELS(τ) (6.15). Therefore, we may wish to recalculate the performance measures periodically as the followup time increases. It is then important to study the updating of performance measures for predictors on a cohort.

Let us consider the probabilistic predictor. Denote t the time when prediction is made and suppose that at time t, the followup time for individual j (i.e. time since their survival time origin) is $c_j(t)$. We then consider IBS(τ)(6.13) and LS(τ) (6.14) at time t with $\tau = c_j(t)$. For example, the truncated logarithm score for individual j at calendar time t is given by $\mathrm{LS}(y_j; c_j(t))$. Then at a subsequent time $t+s$, the updated LS takes the form

$$\mathrm{LS}(y_j; c_j(t+s)) = \begin{cases} \mathrm{LS}(y_j; c_j(t)) & \text{if } y_j \le c_j(t) \\ \mathrm{LS}(y_j; c_j(t)) - \log\left(\dfrac{\hat{f}_p(y_j|z_j)}{\hat{S}_p(c_j(t)|z_j)}\right) & \text{if } y_j \in (c_j(t), c_j(t+s)] \\ \mathrm{LS}(y_j; c_j(t)) - \log\left(\dfrac{\hat{S}_p(c_j(t+s)|z_j)}{\hat{S}_p(c_j(t)|z_j)}\right) & \text{if } y_j > c_j(t+s) \end{cases}$$

Note that once individuals are observed to fail, that is, $y_j \le c_j(t)$, their logarithm scores remain constant. Similarly, we can update $\mathrm{IBS}(y_i, c_j(t))$ as t increases,

$$\begin{aligned} \mathrm{IBS}(y_j, c_j(t)) &= \int_0^{c_j(t)} [I(y_j \le s) - \hat{F}_p(s|z_j)]^2 ds, \\ \mathrm{IBS}(y_j, c_j(t+s)) &= \mathrm{IBS}(y_j, c_j(t)) + \int_{c_j(t)}^{c_j(t+s)} [I(y_j \le s) - \hat{F}_p(s|z_j)]^2 ds. \end{aligned}$$

The above equation suggests that the integrated Brier score may increase as t increases after the observation of individual's survival time, in which case LS stays constant. This form of update is also convenient when there are time-varying covariates. In that case, it is useful to consider the increments to the individual scores at time interval $(t, t+s)$. We intend to develop the methodology for sequential validation and assessment of $\hat{F}_p(y|z)$ when censoring is present.

7.2.4 Additional topics

Many other topics related to prediction in general or specifically for survival times are worth investigation. For example, we can check for the unconditional probabilistic calibration (6.2) with the uniform probability plot of residuals. But the method to check for the stronger version of probabilistic calibration (6.1) has yet to be developed. We are interested to address this issue. In addition, in section 6.1 we mentioned that methods are available to improve the probabilistic calibration of $F_{\hat{\theta}}(y)$ for finite n under conditions i) the family $F_\theta(y)$ we specify includes the true distribution $F_T(y)$; ii) no covariates are allowed. It would be interesting to extend some of the methods to the regression models with covariates so that the calibration of the "plug-in" predictive distribution $F_{\hat{\theta}}(y|z)$ can be improved and consequently a well-calibrated prediction interval for Y given $Z = z$ can be obtained.

Survival analysis is regarded as a special case of event history analysis, which models the occurrence of various types of events. Interesting and challenging prediction problems in event history analysis include the prediction of time to multiple endpoints, the prediction of the number of events in a given time period, updating prediction as events take place, and etc. We would like to generalize the methodologies developed for survival models to the event history framework and provide measures for prediction performance.

The consistency of IPCW estimators depends on the correct specification of the censoring time distribution. Gerds and Schumacher (2006) studied the misspecification of the censoring time distribution and showed that the IPCW estimator $\widehat{\mathrm{BS}}(y)$ could have considerable bias. It is of interest then to compare the bias introduced by misspecified censoring distribution with the bias introduced by discarding censored observations, as some authors did (e.g. Henderson and Keiding 2005).

Bibliography

Adams NM, Hand DJ (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, 32:1139–1147.

Akritas MG (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*, 22:1299–1327.

Bair E, Tibshirani R (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2:511–522.

Brier GW (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Copas J (1999). The effectiveness of risk scores: the logit rank plot. *Applied Statistics*, 48:165–183.

Davison AC, Hinkley DV (1997). Bootstrap methods and their application. Cambridge University Press, Cambridge UK.

Dawid AP (1984). Statistical theory: the prequential approach (with discussion). *Journal of the Royal Statistical Society. Series A*, 147:278–292.

Efron B (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331.

Efron B (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81:461–476.

Efron B (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99:619–632.

Efron B, Tibshirani R (1993). An introduction to the bootstrap. Chapman and Hall, New York.

Efron B, Tibshirani R (1997). Improvement on cross-validation: the 0.632+ bootstrap method. *Journal of the American Statistical Association*, 92:548–560.

Fleming TB, Harrington DP (1991). Counting processes and survival analysis. John Wiley and Sons, New York.

Geisser S (1993). Predictive inference: an introduction. Chapman and Hall, New York.

Gerds TA, Schumacher M (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48:1029–1040.

Gerds TA, Schumacher M (2007). Efron-type measures of prediction error for survival analysis. *Biometrics*, 63:1283–1287.

Gneiting T, Balabdaoui F, Raftery AE (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B*, 69:243–268.

Gould A, Lawless JF (1988). Consistency and efficiency of regression coefficient estimates in location-scale models. *Biometrika*, 73:535–540.

Graf E, Schmoor C, Sauerberei W, Schumacher M (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18:2529–2545.

Guo L, Ma Y, Ward R, Castranova V, Shi X, Qian Y (2006). Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. *Clinical Cancer Research*, 12:3344–3354.

Heagerty P, Lumley T, Pepe MS (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56:337–344.

Heagerty P, Zheng Y (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61:92–105.

Henderson R (1995). Problems and prediction in survival data analysis. *Statistics in Medicine*, 14:161–184.

Henderson R, Jones M, Stare J (2001). Accuracy of point predictions in survival analysis. *Statistics in Medicine*, 20:3083–3096.

Henderson R, Keiding N (2005). Individual survival time prediction using statistical models. *Journal of Medical Ethics*, 31:703–706.

van Houwelingen JC, le Cessie S (1990). Predictive value of statistical models. *Statistics in Medicine*, 9:1303–1325.

Korn EL, Simon R (1990). Measures of explained variation for survival data. *Statistics in Medicine*, 9:487–503.

Korn EL, Simon R (1991). Explained residual variation, explained risk and goodness of fit. *The American Statistician*, 45:201–206.

Krall JM, Uthoff VA, Harley JB (1975). A step-up procedure for selecting variables associated with survival. *Biometrics*, 31:49–57.

Kullback S, Leibler R (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

van der Laan MJ, Hubbard AE, Robins JM (2002). Locally efficient estimation of a multivariate survival function in longitudinal studies. *Journal of the American Statistical Association*, 97:494–507.

Lawless JF (2003). Statistical models and methods for lifetime data, 2nd ed. John Wiley and Sons, Hoboken, NJ.

Lawless JF, Fredette M (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92:529–542.

Lawless JF, Singhal K (1978). Efficient screening of nonnormal regression models. *Biometrics*, 34:318–327.

Li H, Gui J (2004). Partial cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20:i208–i215.

Mallows CL (1973). Some comments on c_p. *Technometrics*, 15:661–675.

Molinaro AM, Simon R, Pfeiffer RM (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21:3301–3307.

Pepe MS (2003). The statistical evaluation of medical tests for classification and prediction. Oxford university press, Oxford.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria, http://www.R-project.org.

Robins JM, Rotnitzky M (1992). Recovery of Information and adjustment for dependent censoring using surrogate markers. In:Jewell N, Dietz K, Farewell V (eds) *AIDS Epidemiology: Methodological Issues*, Birkhauser, Boston, pp 279–331.

Rosthøj S, Keiding N (2004). Explained variation and predictive accuracy in general parametric statistical models: the role of model misspecification. *Lifetime Data Analysis*, 10:461–472.

Schemper M, Henderson R (2000). Predictive accuracy and explained variation in cox regression. *Biometrics*, 56:249–255.

Schemper M, Stare J (1996). Explained variation in survival analysis. *Statistics in Medicine*, 15:1999–2012.

Schumacher M, Graf E, Gerds T (2003). How to assess prognostic models for survival data: a case study in oncology. *Methods Inform. Med.*, 42:564–567.

Shimodaira H (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244.

Soh C, Harrington DP, Zaslavsky AM (2008). Reducing bias in parameter estimates from stepwise regression in proportional hazards regression with right-censored data. *Lifetime Data Analysis*, 14:65–85.

Stein CM (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151.

Taylor JMG, Yu M, Sandler HM (2005). Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology*, 23:816–825.

Therneau TM, Grambsch PM (2000). Modelling survival data: Extending the Cox model. Springer, New York.

Tian L, Cai T, Goetghebeur E, Wei LJ (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika*, 94:297–311.

Tibshirani R (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16:385–395.

Tibshirani R, Knight K (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society. Series B*, 61:529–546.

Uno H, Cai T, Tian L, Wei L (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102:527–537.

White H (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.

Ye J (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131.