Summarising predictive ability of a survival model and applications in medical research

This dissertation is submitted for the degree of
PhD

by

Babak Choodari-Oskooei
July, 2008

University College London &
MRC Clinical Trials Unit

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

I, Babak Choodari-Oskooei, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Summarising predictive ability of a survival model and applications in medical research

by

Babak Choodari-Oskooei

Submitted to the University College London
on July 2008, in partial fulfillment of the
requirements for the degree of
PhD

Abstract

With the molecular revolution in medicine, many new potential prognostic and predictive factors are becoming available. However, whether new factors will lead to substantial improvement in the accuracy of prognostic assessments requires the use of a suitable performance measure when considering different prognostic models. Several such measures have been proposed for use in survival analysis with a particular emphasis on measures proposed for the Cox proportional hazards model. However, there is no consensus of opinion on this issue. The proposed measures make use of a wide spectrum of techniques from information theory to statistical imputation. No comprehensive systematic summary of these measures has been done, and no adequate comparison of measures, theoretically or in practice, has been reported.

This PhD studies the proposed measures systematically. It defines a set of criteria that a measure should possess in the context of survival analysis. Essential aspects of a measure are that it should be consistent under different degrees of censoring and sample size conditions; it should also possess properties such as variable and parameter monotonicity. Desirable properties of a measure are robustness and extendability. This thesis compares the existing measures using these criteria discussing their strengths and shortcomings.

From a practical point of view, a discussion of why these measures are important and what information they can provide in medical research, practical data analysis, and perhaps most importantly in prognostic modelling is presented. Data has been taken from completed randomised controlled trials in several diseases carried out by MRC Clinical Trials Unit and other research organisations. The measures that have the best properties will be applied to models fitted to these datasets. This allows us to quantify and assess the prognostic ability of the available prognostic factors in several diseases.

Thesis supervisor: Patrick Royston
Title: Professor

Thesis Co-supervisors: Mahesh Parmar and Rumana Omar
Title: Professor and PhD

Contents

List of Figures

List of Tables

Acknowledgement 1 *I would like to thank the following people for lending their support and inspiration during the completion of this work:*

Patrick Royston for his supervision and advice, as well as for his patience especially in the early stages of my work. Max Parmar, my other supervisor at the MRC Clinical Trials Unit, for providing many fruitful ideas for this research.

I would also like to thank Dr. Rumana Omar, my other supervisor at the UCL for giving guidance on the organisational side of the PhD.

Last but not least I would like to express my sincere appreciation to my family for their wonderful cooperation, understanding and support throughout the period of this research.

Chapter 1

Introduction

1.1 The context of the research

In the last century, considerable progress was achieved in understanding the aetiology of many diseases. However, both treatment of individual patients and foreknowledge about the outcome of a disease remains a matter of particular importance. In all diseases, there exist factors which assist clinicians in acquiring this knowledge and predicting the prognosis of patients. Such factors are called "prognostic factors". Prognostic factors are useful in a number of ways. Knowledge of prognostic factors can help us understand how the disease would behave if it were untreated, or is likely to behave if treated. Identification of potential prognostic factors may also provide information useful to understand disease mechanisms and help devise new treatments.

One of the objectives in prognostic factor studies is to identify factors that can be used to guide clinical management of patients. To clinicians, knowledge of the relative importance of prognostic factors is invaluable since they usually combine such knowledge with experience to informally help them make decisions about the care of their patients. Laupacis et al (1997) [58] described how clinicians can use prognostic factors to devise clinical rules which assist them in medical decision making when caring for their patients. In general, these rules are created by multivariate regression analysis and either provide the probability of observing an specific outcome, or suggest a diagnostic or therapeutic course of action.

In clinical research, especially in the study of cancer, an understanding of prognostic factors is important in the design and analysis of clinical trials and retrospective reviews

of clinical experience. As Simon (1984) [105] pointed out, it is very difficult to design good clinical trials when prognosis is poorly understood. Valid comparison of treatment and control groups requires that the expected outcome without treatment should be similar in both. Prognostic factors are used as eligibility criteria to ensure a relatively uniform study population, and they may also be used in the process of stratification that is undertaken to balance the case mix in each arm as far as possible.

The necessity to assess the impact of prognostic factors on the survival outcome of patients has given rise to considerable numbers of studies every year. The results of studies are usually summarized in the form of statistics resulting from statistical significance testing, i.e. estimated parameters, confidence intervals, and p-values. Sole dependence on these statistics may lead to misinterpretation of the findings of a study. As Ludwig (2005) [66] stated, statistical tests and p-values give very little information because they can answer only the one very specific question: "Does an observed difference exceed that which might reasonably be expected solely as a result of sampling error and/or random allocation of individuals?" They do not inform us whether prognostic factor information will lead to substantial improvement in the prognostic assessment. There is a great deal of literature about the use and misuse of p-values and statistical tests (Ludwig (2000) [65]; Ludwig (2005) [66]; Igles et al (2001) [73]; Cohen (1994) [16]). Small p-values say nothing about the clinical relevance of the results or the size of the effect. Small p-values can always be obtained with large samples no matter what the true relationship is and how much random experimental error is present. As many, including Abelson (1985) [1], have wisely cautioned, statistical significance tests and p-values should always be used "for guidance rather than for sanctification".

To determine whether research results are of practical significance, we often need to supplement p-values and parameter estimates with statistics that measure the effect magnitude of prognostic factors and new treatments. A variety of statistics have been introduced to measure effect magnitude. Many of the statistics fall into one of two main categories: measures of effect size (typically, standardized mean differences between treatment and control groups) and measures of strength of association (Kirk (2007) [52]). In normal linear regression, R^2 is a standard measure of strength of association between the outcome and predictors. It is also a measure which can be used to further understand the clinical importance of prognostic factors. This measure can help to quantify the improvement in predictive ability when using information on a set of prognostic factors compared to using another set or not using them at all.

1.1.1 Predictive ability in linear regression

The coefficient of determination, R^2, is a well known measure in normal linear regression, which is applied to quantify the predictive ability of covariates, i.e. prognostic factors, in the model. The primary reason for its application is its interpretation as the proportion of variability in the outcome explained by a model, where variability is measured by the variance of outcome variable. In general, the more variability is explained, the better the predictive ability of the model. In other words, R^2 measures how well the model explains the occurrence of different values of the outcome. Furthermore, R^2 quantifies how close the model based predictions are to the observed values of the outcome. It is also a measure of randomness in the outcome that is explained by the model. Kullback and Leibler (1951) [55] applied Shannon's information function [104], which can be used to quantify the amount of information, in statistics and introduced the Kullback-Leibler information gain [55] or divergence measure. They showed that R^2 can be expressed through the Kullback-Leibler distance between models. Due to the link with information gain (i.e. reduction in entropy), R^2 can also be interpreted as the proportion of 'randomness' in the outcome that is explained by the model.

In summary, R^2 is a measure of explained variation and explained randomness, as well as a measure of predictive accuracy for individuals in the study.

1.1.2 Applications of R^2

The coefficient of determination, R^2, has wide applications in medical research and practical data analysis. Some of the applications of R^2 are described below.

To quantify our knowledge of the disease under study

An important application of a predictive ability measure is its use to quantify our knowledge of the disease under study. R^2 as a measure of explained variation can also be considered as a tool to help in finding out how much we know about a disease. In prognostic modelling where the goal is to develop a model which describes the outcome as well as possible, a suitable measure can tell us how much variation in the outcome is explained by prognostic factors in the model. After constructing the prognostic model, the remaining unexplained variation is usually attributed to the random error term in the model. The concept of randomness is an interesting one, as it could be argued that

no events in nature are truly random; we may not know all their influencing factors, and thus they just appear random to our limited knowledge.

For instance, if 20% of the variation in the outcome variable, e.g. survival of patients, is explained by known prognostic factors, this tells us that much remains to be known about the disease. Explaining the remaining proportion of variation may, in theory, be available to a more sophisticated system of prognostic determination, perhaps by using molecular or other types of marker. On the other hand, if 90% of variation in the outcome variable is explained by a model, which is (perhaps) unlikely for many diseases, it tells us that our level of knowledge about the disease is very high.

Effectiveness of surrogate endpoints

A measure like R^2 can be used to evaluate the effectiveness of surrogate endpoints. Herson (1989) [45] wrote that "a surrogate endpoint is one that an investigator deems as correlated with a true endpoint of interest but that can perhaps be measured at lower expense or at an earlier time than the endpoint of interest". Therefore, surrogate endpoints are only useful if they are a good predictor of clinical outcome.

The validation of surrogate endpoints has been studied by Prentice (1989) [83]. He presented a definition as well as a set of criteria, which are equivalent only if the surrogate and true endpoints are binary. Before a surrogate endpoint can replace a final endpoint in the evaluation of an experimental treatment, it must be formally 'validated'. Freedman and Graubard (1992) [28] supplemented these criteria with the so-called 'proportion explained', which is the proportion of the treatment effect mediated by the surrogate. Buyse and Molenberghs (1998) [13] discussed some problems with this class of measures and proposed to replace it with new measures. One of their proposed measures was the individual-level association between the endpoints, after accounting for the effect of treatment, and referred to it as 'adjusted association'. This is one of the applications of R^2 which evaluates the individual level association between the predictor and the outcome. A similar measure can also be used to validate the effectiveness of a surrogate endpoint where the outcome is the survival time.

Practical data analysis

Another application of predictive ability measures is in practical data analysis. In medical research, continuous variables are often converted into categorical variables by grouping values into two or more groups. Royston et al (2006) [90] and Altman and Royston (2006) [5] explained the consequences of converting the continuous data into groups. They also presented alternative methods that make full use of information at hand. A suitable measure of predictive ability can be used to quantify to what extent predictive ability of a continuous variable is diminished, if at all, by recording it as a dichotomy, trichotomy, or more groups. In other words, to what extent do we lose, or gain, by recording a continuous prognostic factor, for example age at diagnosis, into discrete classes on the basis of cutpoints.

Model validation

Measures of predictive ability can also be used for model validation. As Harrell (2001) [36] stated, "model validation is done to ascertain whether predicted values from the model are likely to accurately predict responses on future subjects not used to develop our model". Altman and Royston (2000) [4] examined some general approaches to model validation and discussed two kinds of model validation: internal validation and external validation. Data-splitting, bootstrapping, and cross-validation are methods that can be applied for internal validation of a prognostic model. Measures of predictive ability can be used to evaluate the quality of the predictions obtained from prognostic models. For example, a suitable measure of predictive ability can be used to validate a model in data-splitting methods where we have training and test data sets. Suppose we have training and test samples, each with the same covariates recorded. A model is developed on the training data, its predictive ability, R^2, is estimated, and the model's performance is evaluated on the test data. Royston (2006) [88] demonstrated how this can be done in practice.

1.1.3 Measures of predictive ability in survival models

Survival time studies are an important part of clinical research. There have been serious efforts in the last two decades to devise a measure of predictive ability for statistical models in the analysis of survival data. However, the presence of censoring makes the

definition of such measure much more complicated. Nonetheless, several measures of predictive ability have been proposed for use in survival models, almost exclusively for the Cox proportional hazards regression model. As Schemper and Stare (1996) [99] indicated, there is no simple, easy to calculate, easy to interpret measure for survival models, and in particular for the Cox proportional hazards (PH) regression.

The interpretation of R^2 in linear regression as a measure of explained variation, explained randomness, and predictive accuracy has given rise to a wide spectrum of measures for survival models. We, therefore, classify the proposed measures as measures of explained variation, measures of explained randomness, and measures of predictive accuracy in survival models. We refer to all of them as measures of "predictive ability" in this thesis. The last known attempt to compare the proposed measures was done by Schemper and Stare in 1996 [99]. The outcome of their investigation will be discussed in chapter 3 in more detail. Graf and Schumacher (1995) [32] demonstrated the conceptual differences between some explained variation and predictive accuracy measures in survival models. Furthermore, Henderson et al (2001) [43] investigated the reliability of point predictions derived from familiar survival models by applying some of the measures to real data sets. Several new measures have been proposed since then and there has been no attempt in the literature to compare these measures systematically with regard to a set of criteria. The measures have been mainly studied by the authors who proposed the measures - with the exception of Schemper and Stare (1996) [99]. Despite all the promising properties that were presented by the authors of the measures, the shortcomings of these measures have rarely been addressed. For example, their behaviour has rarely been assessed in the context of multiple regression. Moreover, previous studies lack investigation of these measures systematically across several diseases and real data sets.

This thesis is a study of measures that have been proposed to quantify the predictive ability of covariates in survival models. It investigates their statistical properties and their application in medical research. In addition, it studies the proposed measures across several diseases to quantify and assess the predictive ability of available/known prognostic factors. Great variation exists in the application of measures for examining predictive ability in survival models. Even when investigators use the same measure for a similar population, the estimates of selected measure sometimes differ substantially. Finally, several thorny statistical issues have been raised regarding properties of measures; in particular: variations in formulae, identification and selection of a suitable measure, the effect of censoring, the impact of highly skewed covariates, the relationship between

likelihood and some of the measures, and the maximum value that the measures can reach. We try to address these issues in this thesis with the aim of recommending a small number of measures for general use.

1.2 Organisation and overview

This thesis consists of 9 chapters. This chapter provides an overview. It discusses why measures of predictive ability are important and what information they can provide in medical research and practical data analysis. It also explains their potential use in quantifying our knowledge about a disease.

In chapter 2, first, we give an overview of the measure of predictive ability in linear regression, R^2. Second, an introduction to survival models and proposed measures of predictive ability for survival time data is given. The measures are classified into three main categories and some details on their background are presented. However, there exist measures which use a completely different approach to characterising the predictive ability in survival models such as rank correlation or proportional reduction in log-likelihood. These measures comprise a separate category named as "the other proposed measures". More statistical details of these measures are included in Appendix B.

In chapter 3 we assess the proposed measures with regard to a set of criteria. This chapter formulates our approach and provides us with a framework to study these measures systematically. The measures have been developed based on broad and elusive concepts. The criteria, which are important in the context of survival analysis, simplify the process of drawing conclusions. Tables that summarise these measures according to the proposed criteria will be presented. A thorough investigation of proposed measures in chapter 3, with regard to the proposed criteria, leads to a short-list of measures which might be considered as "potentially recommendable".

In chapter 4, we propose simulation studies to further study the "potentially recommendable measures". We explain the limitations of previous studies and explain the need for further investigation of these measures. This chapter also describes the simulation study design. The simulation studies are mainly designed to investigate the measures with respect to the criteria which are established in chapter 3. Chapter 4 also explains the data generation process and different aspects of the simulation study.

Chapter 5, 6, and 7 study the proposed explained variation measures, explained ran-

domness measures, and predictive accuracy measures, respectively. Each chapter explores the measures in each category through a series of simulation studies and illustrates the results of the study with a set of tables and graphs. In each chapter the impact of censoring and sample size are studied, together with parameter and number of variables monotonicity properties described as criteria in chapter 3. The impact of covariate distribution will be studied by considering distributions with different skewness. The presence of extreme observations in normal linear regression inflates R^2. The R^2 in normal linear regression is also sensitive to outlier observations. In each chapter we assess the impact of extreme and outlier, known an atypical, observations on the measures in that category. Most of the measures are proposed in the context of Cox proportional hazards (PH) regression model. In the presence of non-proportional hazards, the behaviour of these measures is not clear. We discuss this issue together with the impact of model mis-specification.

Chapter 8 is devoted to the application of these measures to medical research and practical data analysis. Data sets from several diseases are considered in this chapter. Finally, chapter 9 presents conclusions of this thesis with some practical recommendations.

Chapter 2

Measures of predictive ability in survival models

2.1 Introduction

Hardin and Hilbe (2007) ([34]) indicated that different interpretations of R^2 in linear regression have given rise to a wide class of measures in nonlinear models. Understanding R^2 in normal linear regression helps us to study many proposed measures in survival models, including their motivation and background. Therefore, in this chapter, first the measure of predictive ability, R^2, in normal linear regression is presented, together with some warning points with respect to its application. Second, an introduction to the Cox proportional hazards (PH) model is given. Then, an introduction to the proposed measures of predictive ability in survival models and their motivation is presented; further statistical details are included in Appendix B for some measures.

Most of the measures are proposed for the Cox PH model. These measures can be classified into three main categories: a) measures of explained variation; b) measures of explained randomness; and c) measures of predictive accuracy. However, other measures of predictive ability proposed for the survival models exist which do not belong to the above categories, such as the proportional reduction in log-likelihood proposed by Harrell (1986) [35] and a measure based on the rank correlation between the imputed survival times and the covariates, proposed by Schemper and Kaider (1997) [98]. We classify them as a completely different category named "other proposed measures" in this thesis.

As presented by Schemper and Stare (1996) [99] and Xu and O'Quigley (1999) [116],

sometimes measures in the same category differ substantially because they measure different population quantities. Nonetheless, the proposed classification helps us to grasp the theoretical underpinning of the measures and facilitates their interpretation. In normal linear regression, R^2 measures all three of explained variation, explained randomness/uncertainty, and predictive accuracy. Outside the linear regression model, the measures usually differ. Understanding this distinction is essential in order to draw correct conclusions in practice.

2.1.1 R^2 and its interpretation in simple linear regression

Let $X_1, X_2, ..., X_p$ and Y denote $p+1$ random variables. In the standard linear regression model $X_1, X_2, ..., X_p$ typically denote independent variables or covariates, usually called predictors or explanatory variables, and Y typically denotes the dependent variable, also known as the outcome variable. The regression function is linear and the model can be stated as

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + e \tag{2.1}$$

where $Y = (y_1, ..., y_n)'$, $X = (x_1, x_2, ..., x_p)$ is a fixed $n \times p$ design matrix, β_0 and β_j $(j = 1, ..., p)$ represent the unknown parameters, $e = (e_1, ..., e_n)'$ is a vector of independent errors with $E(e_i) = 0$, $var(e_i) = \sigma^2$ $(i = 1, ..., n)$, and n is the total sample size.

Let us assume the simplest model with one dependent and one independent variable where we have Y and one X. Figure 2-1 shows the observation Y_i for the values of X_i. The variation in Y_i is conventionally measured in terms of the deviation of $Y_i s$ around their mean $\overline{Y}$, i.e. $Y_i - \overline{Y}$, which is specified by a vertical line for observation i in figure 2-1. The measure of total variation, denoted by SST, is the sum of the squared deviations, $SST = \sum(Y_i - \overline{Y})^2$. The greater the variation among Y_i observations, the higher the value of SST.

When we use the predictor variable X, the variation reflecting uncertainty in the outcome variable Y is $Y_i - \widehat{Y}_i$ that of the Y_i observations around the fitted regression line. A measure of variation in the Y_i when regression on the predictor variable X is taken into account is the sum of the squared deviations, $SSE = \sum(Y_i - \widehat{Y}_i)^2$. The greater the variation of Y_i observations around the fitted regression line, the higher the value of SSE.

The difference between SST and SSE accounts for the regression sum of squares

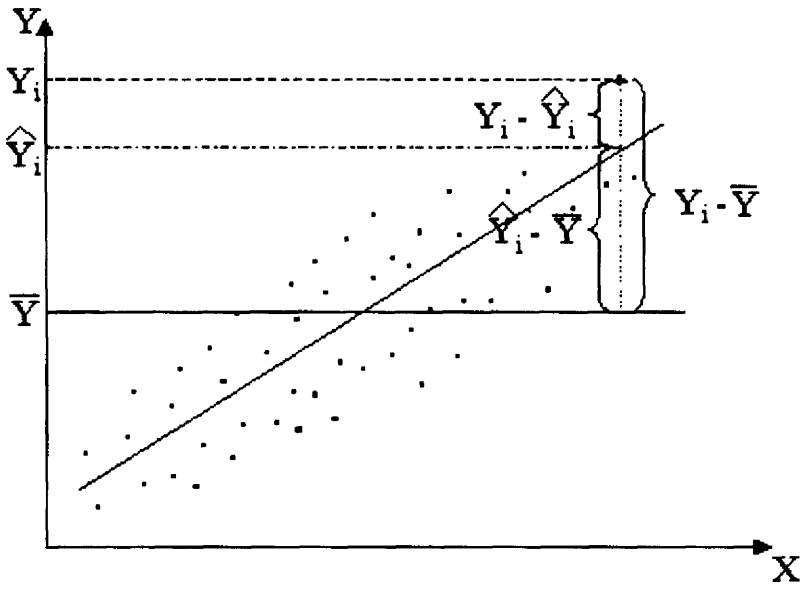


Figure 2-1: Relationship between Y and estimated Y in simple linear regression

$SSR = \sum(\hat{Y}_i - \overline{Y})^2$. The total deviation $Y_i - \overline{Y}$ can be decomposed into two components.

$$\underbrace{Y_i - \overline{Y}}_{\textit{Total deviation}} = \underbrace{\hat{Y}_i - \overline{Y}}_{\textit{Dev. of fitted regression value from mean}} + \underbrace{Y_i - \hat{Y}_i}_{\textit{Dev. of data from fitted regression line}} \tag{2.2}$$

Figure 2-1 shows the decomposition for observation Y_i by dotted line. It can be shown that the sums of these squared deviations have the same relationship, i.e. the total sum of squares, SST, is equal to the sum of regression sum of squares, SSR, and error sum of squares, SSE.

$$\begin{aligned}
\sum (Y_i - \overline{Y})^2 &= \sum \left[\left(\hat{Y}_i - \overline{Y}\right) + \left(Y_i - \hat{Y}_i\right) \right]^2 \\
&= \sum \left[\left(\hat{Y}_i - \overline{Y}\right)^2 + \left(Y_i - \hat{Y}_i\right)^2 + 2\left(\hat{Y}_i - \overline{Y}\right)\left(Y_i - \hat{Y}_i\right) \right] \\
&= \sum \left(\hat{Y}_i - \overline{Y}\right)^2 + \sum \left(Y_i - \hat{Y}_i\right)^2 + 2\sum \left(\hat{Y}_i - \overline{Y}\right)\left(Y_i - \hat{Y}_i\right)
\end{aligned}$$

The last term on the right equals zero, as we can see by expanding it:

$$2\sum \left(\hat{Y}_i - \overline{Y}\right)\left(Y_i - \hat{Y}_i\right) = 2\sum \hat{Y}_i \left(Y_i - \hat{Y}_i\right) - 2\overline{Y}\sum \left(Y_i - \hat{Y}_i\right).$$

The first summation on the right is zero, which is one of the properties of fitted

regression line, and the second equals zero since it is the sum of the residuals. SST measures the variation in the observation Y_i, or the uncertainty in predicting Y, when the predictor variable X is not taken into account. Similarly, SSE measures the variation in the Y_i when a regression model utilizing the predictor variable X is considered. A natural measure of the effect of X in reducing the variation in Y, i.e. in reducing the uncertainty in predicting Y, is to state the decrease in variation ($SST - SSE = SSR$) as a proportion of the total variation:

$$R^2 = \frac{\textit{variation in } y \textit{ explained by regression}}{\textit{total observed variation in } y} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{2.3}$$

The measure R^2 is called the *coefficient of determination.* Since $0 \leq SSE \leq SST$, it follows that $0 \leq R^2 \leq 1$.

Several formulae have been presented for R^2 for the linear regression in the literature, which lead to the same results in simple linear regression. Kvalseth (1985) [56] listed some of them as follows.

$$R_1^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{SSE}{SST} \tag{2.4}$$

$$R_2^2 = \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = \frac{SSR}{SST} \tag{2.5}$$

$$R_3^2 = \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{\widehat{Y}})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} \tag{2.6}$$

$$R_4^2 = 1 - \frac{\sum_{i=1}^n (e_i - \overline{e})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} \tag{2.7}$$

where $e_i = Y_i - \overline{Y}$ and $\overline{e} = \frac{1}{n} \sum_{i=1}^n e_i$.

$$\begin{aligned} R_5^2 &= \textit{squared multiple correlation coefficient between response variable and predictor} \\ &= r^2(Y_i, X_i) \end{aligned} \tag{2.8}$$

$$\begin{aligned} R_6^2 &= \textit{squared correlation coefficient between } Y_i \textit{ and } \widehat{Y}_i \\ &= r^2(Y_i, \widehat{Y}_i) \end{aligned} \tag{2.9}$$

$$R_7^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^n Y_i^2} \tag{2.10}$$

$$R_8^2 = \frac{\sum_{i=1}^{n} \widehat{Y}_i^2}{\sum_{i=1}^{n} Y_i^2} \tag{2.11}$$

However, care must be exercised as the above definitions of R^2 lead to different results in models where suitable transformations of the variables are used to form standard linear models (Kvalseth 1985 [56]). Besides, they result in different values in models without intercept or when different methods of estimation other than linear least squares are used (see section 2.1.2). In these cases, values of R^2 derived from the above definitions would not necessarily be the same.

2.1.2 Health warnings against model comparisons using R^2

Despite its usefulness and common application in linear regression, R^2 may easily be misinterpreted by research workers. For example, R^2 can be misinterpreted as a measure of goodness of fit. Magee (1990) [68] and Vandaele (1981) [111] showed the relationship between R^2 and measures of model fit. These relationships were defined to suggest logical extensions of R^2 and show problems that may arise in its application. However, R^2 is an inappropriate measure to compare the fit of competing regression models for which the underlying null models are not identical. To explain this in more detail, consider the general definition of R^2 in (2.3). R^2 measures only how much the model (2.1) improves the null model, i.e. a model with just an intercept. R^2 is the proportion of variation in the outcome variable that can be accounted for by incorporating a covariate into a particular model instead of viewing the outcome variable by itself. An important feature of R^2 is that it is used to compare models for which the underlying null model is identical, e.g. nested models.

Royston (2006) [88] showed how ignoring this characteristic of R^2 may cause data analysts to reach misleading conclusions in practice. He compared measures of goodness of fit and R^2s of a range of survival models including gamma, Weibull, and lognormal. The results of his analysis showed that the R^2 for the gamma model is lower than that of the Weibull or lognormal models. The gamma model has the lowest deviance, i.e. $-2l_{\widehat{\beta}}$, for the null model and the model with covariates because it fits the underlying distribution better than the other two models, leaving less scope than the other models for the inclusion of covariates to improve the fit . Nevertheless, as judged by the AIC,

$$AIC = -2(\text{log likelihood}) + 2(c + p + 1)$$

where c is the number of covariates and p is the number of model-specific ancillary parameters, the Weibull and lognormal models with covariates fit worse than the gamma model. Therefore, the common belief that large R^2 demonstrates model adequacy or model superiority is proven to be wrong.

Anderson-Sprecher (1994) [6] and Scott and Wild (1991) [103] presented some of the strongest warnings about the mis-application of R^2 in model selection when the values of the coefficient are calculated in different contexts. Some of the points that should be considered in the application of R^2 in model building are as follows.

Transformation and R^2

To identify the correct functional form between the outcome and the predictors, it is often advantageous to transform the response variable when a least squares regression model is fitted to a set of data. This can lead to difficulties in making comparisons between competing transformations. Kvalseth (1985) [56] warned of the problems that arise when R^2 is used to compare models that involve different transformations of the response variable. Scott and Wild (1991) [103] reiterated this warning by applying it to real data.

Their example consisted of data on the length of the liver as the response variable and gestational age as a single predictor. Scott and Wild (1991) used two different transformations of the response variable, i.e. logarithmic and power transformations. The results of the study showed that the two models were essentially interchangeable for all practical purposes. Almost all model diagnostic tests including residual tests, predictions, and the fitted curves resulting from the models were identical. The R^2 of the two models was calculated as the squared multiple correlation coefficient between the response variable and the predictor (R_5^2 in equation (2.8)). The R^2s of the two models differed enormously, being 0.13 and 0.88. The exact reason for the big difference in R^2 is beyond the scope of this thesis. To put it in a nutshell, it is the result of a change in the metric of the response variable.

Huang and Draper (2003) [46] examined this problem and gave a thorough explanation in a new way by considering the underlying regression geometry. Greenland (1996) [33] also explained how transformation of the response variable can have profound effects on the correlation coefficient in the lognormal distribution. In summary, R^2 should not be used to compare models with different transformations of the outcome variable.

Models with unequal measures of variation

Different measures of variation in numerator and the denominator of the R^2 definition in (2.3) result in different R^2s. For example, the least median squares estimator (applied in robust regression) uses $med_i(Y_i - \widehat{Y}_i)^2$ as the measure of variation and replacing this quantity in the general definition of the R^2 in (2.3) will lead to

$$R_R^2 = 1 - \frac{med_i(Y_i - \widehat{Y}_i)^2}{med_i(Y_i - M)^2} \tag{2.12}$$

where M is a constant that minimises $med_i(Y_i - M)^2$. Rousseeuw and Leroy (1987) [87] presented R_R^2 for robust regression.

Using different measures of variation will lead to different R^2 values. Examples of methods whose R^2 should not be compared against each other, or against least squares regression, are ridge regression, robust regression, and weighted least squares. In brief, R^2 should not be used to evaluate models that are based on different measures of variation.

In summary, research workers should take due consideration and should definitely be careful in the interpretation of R^2. The R^2 should not be used to compare predictive ability of different models whose null model are different, or models that use different outcome transformation. However, it can be used to compare the predictive ability of nested models.

2.1.3 Adjusted R^2

In theory, using an unlimited number of independent variables to explain the change in a dependent variable would result in an R^2 of 1. Consider the general definition of R^2 in (2.3): SST is fixed (unchanging) and SSR can only increase by adding new independent variables. Therefore, each additional variable used in the regression model will not decrease the SST and will probably increase SSR at least slightly, resulting in a higher R^2. This happens even when the new variable causes the regression model to become less efficient by adding to the variance of the predictions. Ezekiel (1930) [24] proposed an adjusted R^2 that is obtained by dividing two quantities, SSE and SST, by the respective degrees of freedom.

$$R_{adj}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - (1 - R^2)\frac{(n-1)}{(n-p)} \tag{2.13}$$

where n is the sample size and p is the number of independent variables. In addition, Kendall and Stuart (1979) ([48], page 362) showed that R^2 in (2.3) is a biased estimator of the population $\mathbf{R}^2$, which is defined as

$$\mathbf{R}^2 = 1 - \frac{E\{\mathbf{Var(Y|X)}\}}{\mathbf{Var(Y)}}$$

where $\mathbf{Var}(Y|X) = E\left\{[Y - E(Y|X)]^2 |X\right\}$ is the variance of the response around its true regression. Furthermore, Kendall and Stuart (1979) [48] showed that when $\mathbf{R}^2 = 0$

$$E(R^2|\mathbf{R}^2 = 0) = 1 - E\left\{\frac{SSE}{SST}\right\} = \frac{p}{n-1},$$

but

$$E(R^2_{adj}|\mathbf{R}^2 = 0) = 0.$$

So R^2_{adj} is an unbiased estimator when $\mathbf{R}^2 = 0$.

2.2 Survival Models

2.2.1 The proportional hazards model

Let T be a non-negative random variable representing time to an event of interest, e.g. death or disease recurrence, in individuals or objects in the population under study. We will assume T to be continuous. The survival function is defined as $S(t) = \Pr(T > t)$ where Pr denotes the probability. The distribution function of the random variable T, $F(t)$, and survival function, $S(t)$, have the relationship $S(t) = 1 - F(t)$. An important concept in the study of survival time distribution and modelling is the hazard function, which is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t}.$$

The hazard function can be interpreted as a measure of proneness to failure in the interval $[t, t + \Delta t]$, for small Δt, provided that the event has not occurred before t. In contrast to the survival function, which describes the probability of not failing before time t, the hazard function focuses on the propensity to fail at time t among those individuals who have not experienced the event by t. It can be shown that $S(t)$ and $h(t)$ have the following relationship:

$$S(t) = \exp(-\int_0^t h(u)du) \tag{2.14}$$

Cox (1972) [19] introduced the proportional hazards regression model in which the hazard function, $h(t, X)$, is modelled for an individual with covariate vector X, using

$$h(t, X) = h_0(t)\exp(\beta X) \tag{2.15}$$

where $h_0(t)$ is the baseline hazard function. The Cox proportional hazards model does not make any assumption about the shape of the underlying hazard function by using partial likelihood to estimate the underlying regression coefficients, β. In proportional hazards regression, the baseline hazard, $h_0(t)$, vanishes from the partial likelihood, and we obtain only estimates of the regression coefficients associated with the explanatory covariates. The only assumption in the Cox model is that the hazards are proportional.

2.2.2 Generalised survival models

Royston and Parmar (2002) [92] developed flexible parametric models based on the assumption of proportional hazards or proportional odds scaling of the covariate effects or probit. Their proposed class of models are based on the transformation of the survival function by a link function $g(.)$ given by

$$g[S(t; Z)] = g[S_0(t)] + \beta X \tag{2.16}$$

where $S_0(t) = S(t; 0)$ is the baseline survival function and β is a vector of parameters to be estimated for covariates X. They developed three such types of models for survival analysis. They are obtained with the probit, logit and complementary log-log link functions respectively,

$$g(s) = \Phi^{-1}(1 - s) \tag{2.17}$$

$$g(s) = \ln(\frac{1-s}{s}) \tag{2.18}$$

$$g(s) = \ln[-\ln(s)] \tag{2.19}$$

These three link functions generate the regression models

$$\Phi^{-1}[1 - S(t; X)] = \Phi^{-1}[1 - S_0(t)] + \beta X \qquad \text{(probit model)} \tag{2.20}$$

$$\ln(\frac{1-S(t;X)}{S(t;X)}) = \ln(\frac{1-S_0(t;X)}{S_0(t;X)}) + \beta X \quad \text{(proportional odds model)} \tag{2.21}$$

$$\ln[-\ln S(t;X)] = \ln[-\ln S_0(t)] + \beta X \quad \text{(proportional hazards model).} \tag{2.22}$$

They called this class of models generalised survival models or GSMs. Royston and Sauerbrei (2004) [93] applied GSMs to define an R^2-type measure based on discrimination measures for survival data which will be discussed later in this chapter.

2.3 Measures of predictive ability in survival models

This section presents statistical details of three classes of measures of predictive ability in survival analysis. As a first step toward systematic comparisons among predictive ability measures in survival models, some of the proposed measures are computed in this section for illustration using real data. This example helps us to clarify the differences between three main classes of predictive ability measures and to show where they are applicable. The technical details of the measures in each category follows this example.

We work with the breast cancer data set which was analysed in detail by Sauerbrei and Royston (1999) [94]. Further analysis of this data set, along with real data sets from other diseases, will be presented in chapter 8. The data relate to a set of 686 patients with node-positive breast cancer. The outcome of interest is the recurrence-free survival time (RFS), that is the duration in years from entry into the study (typically, the time of diagnosis of primary breast cancer) until either death or disease recurrence, whichever occurred first. There were 299 events for this outcome and the median follow-up time was about 5 years.

Model III of [94] was a Cox proportional hazards model for RFS which included 5 covariates: age with a fractional polynomial transformation with powers -2 and -0.5, tumour grade 2/3, number of positive lymph nodes (PLN) with the exponential transformation $\exp(-0.12 * PLN)$, progesterone receptor with a fractional polynomial transformation with power 0.5, and hormonal therapy with tamoxifen (yes/no). Tables 2.1 to 2.3 show estimated values of some predictive ability measures for model III.

As it is evident from tables 2.1 to 2.3, the values of these measures vary widely, even though all measures are constrained to the $[0, 1]$ range. The selected explained variation measures vary from 0.24 to 0.29. They generally measure the variation in the outcome variable in the model that is 'explained' through the prognostic factors in

Table 2.1: Estimates of some explained variation measures using model III for breast cancer data in Royston and Sauerbrei (1999).

Measures of Explained Variation	Measure Value
Helland (1987) [41], Kent & O'Quigley (1988) [49] - R^2_{PM}	0.27
Royston & Sauerbrei measure (2004) [93] - R^2_D	0.28
Korn & Simon measure (1990) [53] - R^2_{KS}	0.24[1]
Royston measure (2006) [88] - $R^2_{Royston}$	0.29

1: squared error loss was used to evaluate the measure

the model. Therefore, it can be concluded that the available prognostic factors explain about $24\% - 29\%$ of the variation in the outcome variable, whereas the selected explained randomness measures in table 2.2 vary from 0.20 to 0.40. These measures involve the calculation of expected information gain. Because of the link with information gain (i.e. reduction in entropy or randomness as explained in section 2.3.2), Kent & O'Quigley (1988) [49] describe these types of measures as the proportion of 'explained randomness' of a model, rather than explained variation. The selected measures in the third category, predictive accuracy measures, in table 2.3 vary from 0.16 to 0.18. These measures evaluate the individual survival probability predictions from the model. The results in table 2.3 show that providing informative prognosis at the individual level is limited for breast cancer patients since the predictive accuracy that can be achieved with the available prognostic factors is only $16\% - 18\%$.

Table 2.2: Estimates of some explained randomness measures using model III for breast cancer data in Royston and Sauerbrei (1999).

Measures of Explained Randomness	Measure Value
Kent and O'Quigley measure (1988) [49] - ρ^2_W	0.36
Approximation to Kent and O'Quigley [49] - $\rho^2_{W,A}$	0.38
Nagelkerke measure (1991) [71] - ρ^2_n	0.20
Xu & O'Quigley measure (1999) [116] - ρ^2_{XuOQ}	0.37
O'Quigley et al measure (2005) [80] - ρ^2_k	0.40

For the normal-errors regression model without censoring, explained variation, ex-

Table 2.3: Estimates of some predictive accuracy measures using model III for breast cancer data in Royston and Sauerbrei (1999).

Measures of Predictive Accuracy	Measure Value
Schemper's V_1 and V_2 measures (1990) [95]	$V_1 = 0.16$; $V_2 = 0.17$
Graf et al measure [31] - R^2_G	0.16[1]
Schemper and Henderson measure (2000) [97] - V_{SH}	0.18

1: evaluated at the 50th centile of observed survival time

plained randomness, and predictive accuracy (and the resulting statistics) coincide, but for survival models with or without censoring, these statistics are different. The rest of this chapter presents a theoretical summary of the three main classes of predictive ability measures in survival analysis.

2.3.1 Measures of explained variation

The first category contains explained variation measures. The most popular interpretation of R^2 is the percent variance in the outcome that is explained by the covariates. Measures in this category are proposed by Helland (1987) [41], modified to use for the Cox PH model by Kent and O'Quigley (1988), Korn and Simon (1990) [53], Akazawa (1997) [2], O'Quigley and Flandre (1994) [75], O'Quigley and Xu (2001) [78] [79], Royston and Sauerbrei (2004) [93], and Royston (2006) [88]. The measures summarise the proportion of variability in the outcome explained by the model, where variability is measured by a variation function. In general, the more variability explained, the better the predictive ability of the model. The main difference in the proposed measures in this category is in their variation function.

Helland (1987) and Kent and O'Quigley (1988) measure - R^2_{PM}

Helland (1987) [41] proposed an explained variation measure for the linear regression models. He suggested that the population multiple correlation coefficient can be defined as the correlation between the outcome and the linear predictor, i.e. prognostic index $\beta' x$. He concluded that the total variation in the outcome splits into two components: that explained by the covariates, and the remaining unexplained variation. Therefore, an explained variation measure can be defined as the ratio of variation explained by

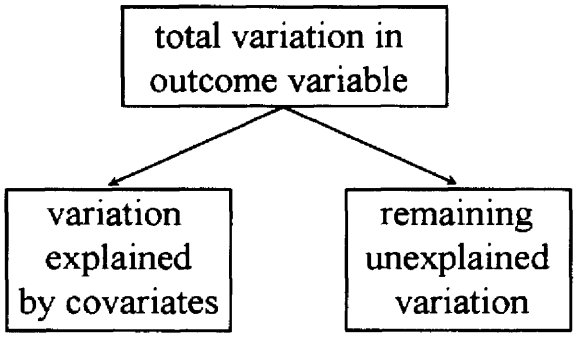


Figure 2-2: Schematic illustration of explained variation measures; the total variation in outcome is divided into two components

covariates in the model to the total variation which consists of two components.

$$R^2_{Helland} = \frac{\mathbf{Var}_X(\beta' x)}{\mathbf{Var}_X(\beta' x) + \sigma^2_\varepsilon} \tag{2.23}$$

where $\mathbf{Var}(\mathbf{Y}) = \mathbf{Var}_X(\beta' x) + \sigma^2_\varepsilon$ is the total variation in the regression model.

This can be generalised for other regression models as

$$E[Y|X] = \beta' X + \varepsilon$$

where ε is the error term. Variance of Y given X is

$$Var(Y|X) = E[Y^2|X] - (E[Y|X])^2$$

and since $E[g(Y)] = E[E[g(Y)|X]]$, specifically $E[Y] = E[E[Y|X]]$, we can write

$$\begin{aligned} E(Var(Y|X)) &= E[Y^2] - (E[Y])^2 - E[(E[Y|X])^2] + (E[Y])^2 \\ &= Var(Y) - E[(E[Y|X])^2] + (E[E[Y|X]])^2 \\ &= Var(Y) - Var(E[Y|X]) \end{aligned}$$

$$
\begin{aligned}
R^2 &= 1 - \frac{E(Var(Y|X))}{Var(Y)} \\
&= \frac{Var(E[Y|X])}{Var(Y)} \\
&= \frac{Var(\beta' X)}{Var(\beta' X) + \sigma_\varepsilon^2}
\end{aligned}
$$

Motivated by the above relationship and Helland's measure (1987) for linear regression models, Kent and O'Quigley (1988) [49] proposed a similar measure, R^2_{PM}, for the Cox PH model.

$$
R^2_{PM} = \frac{\mathbf{Var}_X(\beta' x)}{\mathbf{Var}_X(\beta' x) + \pi^2/6} \tag{2.24}
$$

where $\frac{\pi^2}{6} \simeq 1.645$ is the variance of error term in an equivalent Weibull model. In this model the conditional distribution of T given X is modelled by

$$
Y = \log(T) = -\sigma(\mu - \beta X) + \sigma\varepsilon
$$

where ε is independent of X and has density $f(y)$ where

$$
f(y) = e^y \exp(-e^y),
$$

i.e. the extreme value density. In this case we know that $T = e^Y$ follows a Weibull distribution conditional upon $X = x$ [59].

Royston and Sauerbrei measure (2004) - R^2_D

Royston and Sauerbrei (2004) [93] proposed a measure of explained variation based on the above measure, R^2_{PM}. One of the interests in the survival analysis is in identifying subgroups of patients with different risks of failure. The aim is to define groups which are well-separated and sufficiently substantial to be useful in clinical settings. Royston and Sauerbrei (2004) [93] proposed a measure, D, to assess prognostic separation of survival curves. They applied the D measure to the explained variation measure defined by Helland (1987) and Kent and O'Quigley (1988), i.e. R^2_{PM}, to propose a new measure which is based on the separation of survival curves:

$$R_D^2 = \frac{D^2/\kappa^2}{D^2/\kappa^2 + \sigma_\varepsilon^2} \tag{2.25}$$

where $\kappa = \sqrt{8/\pi}$ and σ_ε^2 is the variance of the error term in the model, where

$$\sigma_\varepsilon^2 = \begin{cases} 1 & \text{(lognormal model or models with probit link)} \\ \pi^2/3 & \text{(log-logistic model or proportional odds model)} \\ \pi^2/6 & \text{(proportional hazards models)} \end{cases}$$

R_D^2 can be used for a wide class of models, including the Cox PH model.

Korn and Simon class of measures (1990)

Korn and Simon (1990) [53] proposed a class of explained variation measures which requires the specification of a loss function, $L(t, \widehat{t})$, that gives the loss incurred from a prediction, $\widehat{t}$, to the observed survival time, t. Their approach leads to a wide range of measures of explained variation depending on the loss function applied to minimise the expected loss. Two common possibilities are absolute error loss, $L(t, p) = |t - \widehat{t}|$, and squared error loss, $L(t, \widehat{t}) = (t - \widehat{t})^2$.

Korn and Simon measures (1990) require the specification of a time range of interest. For example, their loss function approach can quantify the predictive ability of a set of covariates up to 5 years after diagnosis. They suggest using the average of predicted survivals in the denominator of the measures instead of squared error loss given the null model. Henderson (1995) [42] further developed Korn and Simon's (1990) approach by proposing more flexible loss functions. The expected loss (risk) for any loss function, $L(t, \widehat{t})$, is defined as

$$R(x) = \int_0^\infty L(t, \widehat{t}) dF(t|x)$$

where $F(t|x) = 1 - S(t|x)$. For example, $\widehat{t} = E(T|x)$ is the optimal predictor that minimises the expected risk with squared error loss $R = \min_p \int (t - \widehat{t})^2 dF(t|x)$, which is the variance of T. The risk under the null model is defined as $R_0 = \int L(t, \widehat{t}_0) dF_0(t)$ where $\widehat{t}_0$ minimises the expected loss with respect to $F_0(t)$. Then, the explained variation is defined as the proportional decrease in risk obtained by using the covariates in the model.

$$\textit{explained variation} = \frac{R_0 - E_X[R(X)]}{R_0} \tag{2.26}$$

where $E[R(X)]$ is the expected value of $R(x)$ averaged over the distribution of the Xs, i.e. covariate(s). For the censored survival data, Korn and Simon (1990) proposed alternative $R(x)$ which is loss function with squared error loss censored at T_0.

Akazawa measure (1997) -R^2_{Ak}

Akazawa (1997) [2] proposed a similar measure to Korn and Simon's (1990) class of measures. His approach was motivated by the definition of R^2 in normal linear regression in equation (2.6) and is defined as:

$$R^2_{Ak} = \frac{\sum_{i=1}^{n} \left(E[T_i|X_i] - \frac{1}{n}\sum_{i=1}^{n} E[T_i|X_i]\right)^2}{\sum_{i=1}^{n}(t_i - \frac{1}{n}\sum_{i=1}^{n} t_i)^2}. \tag{2.27}$$

where $E[T_i|X_i] = \int_0^{T_0} T dF(T|X_i, T_0)$.

O'Quigley and Flandre measure (1994) - R^2_{OQF}

O'Quigley and Flandre (1994) [75] suggested a measure for the Cox PH model that compares mean squared Schoenfeld residuals [101] under a proportional hazards model to that of the null model. The R^2 in normal linear regression can be defined in terms of prediction errors or residuals, equation (2.5). This measure applies this principle to the Cox PH model, but it considers Schoenfeld residuals. O'Quigley and Flandre (1994) [75] argued that since the Cox semiparametric model leaves inference depending only on the failure time rankings, and being able to predict the failure rankings of all failed subjects is equivalent to being able to predict at each failure time which subject is to fail, it is sensible to measure the discrepancy between the observed covariate at a given time and its expected value under the model. This measure quantifies the predictability of a covariate from a given failure time and is given by:

$$R^2_{OQF}(\beta) = \frac{\sum_{failures\ t_j} r_j^2(0) - \sum_{failures\ t_j} r_j^2(\beta)}{\sum_{failures\ t_j} r_j^2(0)} \tag{2.28}$$

where $r(\beta)$ and $r(0)$ are Schoenfeld residuals [101] under the full and null models, respectively. In the absence of censoring, the quantity $\sum_{i=1}^{n} r_i^2(\beta)/n$ is a residual sum of squares, analogous to SSE in linear regression, and can be viewed as the average discrepancy between the observed covariate and its expectation under the model, whereas $\sum_{i=1}^{n} r_i^2(0)/n$ is the total sum of squares, analogous to SST in linear regression. The

population value of this measure for a single covariate is

$$\mathbf{R}^2_{OQF} = \frac{\mathbf{Var}(X) - \int_0^\infty \mathbf{Var}(X|t)f(t)dt}{\mathbf{Var}(X)} \tag{2.29}$$

where $\mathbf{Var}(X) - \int_0^\infty \mathbf{Var}(X|t)f(t)dt = \mathbf{Var}\left(\mathbf{E}(X|t)\right)$.

Xu and O'Quigley measure (2001) - R^2_{XuOQ}

Xu and O'Quigley (2001) [78] further developed O'Quigley and Flandre's measure (1994) to eliminate any dependence of R^2_{OQF} upon censoring. They did this by weighting the squared Schoenfeld residuals by the increments of consistent estimate of marginal failure time distribution function. Their measure is defined as

$$R^2_{XuOQ}(\beta) = \frac{\sum_{failures\ t_j} W(t_j)r^2_j(0) - \sum_{failures\ t_j} W(t_j)r^2_j(\beta)}{\sum_{failures\ t_j} W(t_j)r^2_j(0)}$$

where $W(t_j)$ is the jump of the Kaplan-Meier curve at an event time t_j.

Measures proposed by O'Quigley and Flandre (1994) and Xu and O'Quigley (2001) [78] exploit partial likelihood estimation method because it provides model-based estimates of the distribution of covariates conditional on survival time. Focusing on a scalar covariate, Xu and O'Quigley (2000) presented an estimate of the distribution of the covariate, X_i, $i = 1, ..., n$, conditional on the event occurring at time t_j, $j = 1, ..., k$, where n and k are the number of individuals and number of events, respectively.

Royston measure (2006) - $R^2_{Royston}$

Finally, Royston (2006) [88] suggested a measure which is a modified version of a measure proposed by O'Quigley et al (2005) [80].

$$R^2_{Royston} = \frac{\rho^2_k}{\rho^2_k + (\pi^2/6)(1 - \rho^2_k)}$$

where ρ^2_k is a measure of explained randomness, presented in the next section, proposed by O'Quigley et al (2005) [80].

2.3.2 Measures of explained randomness

The second category contains explained randomness measures. These measures are based on the notion of information in information theory. Information has a technical meaning, not radically different from the everyday meaning, which is "a numerical quantity that measures the uncertainty in outcome of an experiment to be performed" [107]. In other words, we can gain information only about matters in which we are to some degree ignorant, or uncertain: indeed, information may be defined as that which removes or reduces uncertainty. The important implication of this definition is that once we are able to measure uncertainty, we can also measure information in similar terms.

Several methods have been introduced to quantify the amount of information in the context of communication engineering and information theory [38] [104]. Later, these methods were applied to statistical theory by discussing the notion of information in an experiment. One purpose of experimentation is to reach decisions, another purpose is to gain knowledge about the state of nature, e.g. about parameters in the model. The knowledge is measured by the amount of information, as described below. In this section, we first give a brief overview of information and then present the background to the proposed measures of explained randomness/uncertainty before introducing them in the last subsection.

Information functions

Scientists in information and communication theory have devised methods to express information numerically in the same way as distance, time, mass, temperature, etc. Hartley (1928) [38] introduced the first information function. He stated that the answer to a question that can assume the two values 'yes' or 'no' contains one unit of information, that is one bit. Hartley's formula to measure the amount of information in a set E which contains N elements is

$$I(E) = \log_2(N).$$

For example, suppose that we toss a symmetric coin then the information content of the event 1) having a head, or 2) having a tail is $\log_2(2) = 1$ unit of information, a bit. Later, Shannon (1948) [104] further developed Hartley's formula for sets or elements that do not occur with equal probabilities. Shannon's (1948) function was primarily proposed to quantify the expected uncertainty associated with an outcome from a set of symbols

$\{x_j; j = 1, ..., J\}$ that are received from a source X according to probability distribution $\Pr(X)$. Suppose X is a random variable with possible values $E = \{x_1, x_2, ..., x_N\}$. Let us denote P_k the probability that X assumes the value x_k, $k = 1, 2, ..., N$. Shannon (1948) [104] proposed the following function to calculate the amount of information:

$$\begin{aligned} I(E) &= P_1 \log_2 \frac{1}{P_1} + P_2 \log_2 \frac{1}{P_2} + ... + P_k \log_2 \frac{1}{P_k} \\ &= -\sum_{i=1}^{k} P_i \log_2(P_i) \end{aligned} \tag{2.30}$$

If $P_1 = P_2 = ... = P_k = \frac{1}{N}$ then Shannon's formula reduces to that of Hartley (1928). He called this function "entropy" because of the similarities of his proposed function with the thermodynamic entropy expression. If X is a continuous random variable, the probability distribution, $\Pr(X)$, and summation notation in equation 2.30 are replaced with the density function, $f(x)$, and integral, respectively.

Kullback-Leibler information gain

Kullback and Leibler (1951) [55] applied Shannon's information function to statistics and introduced the Kullback-Leibler information gain [55] or divergence measure. Let H_0 and H_1 be null and alternative hypotheses for a random variable Y defined on a sample space χ with true density $f(y; \alpha_0)$. Under H_0, Y is assumed to follow density $f(y; \alpha_o)$ and, under H_1, it is assumed to follow density $f(y; \beta)$. Sometimes, we shall want to suppose that H_0 is nested in H_1. We regard H_0 as the true model with true parameter α_0. Following the Shannon's formula in $I(E)$ above, $I(\alpha_0|\alpha_0) = \int_y \log\{f(y; \alpha_0)\} f(y; \alpha_0) dy$ is defined as the expected information on α_0 under $f(y; \alpha_0)$, i.e. information at the true parameter value. Similarly, the expected information attached to the value β when the distribution is $f(y; \alpha_0)$ is $I(\beta|\alpha_0, y) = \int_y \log\{f(y; \beta)\} f(y; \alpha_0) dy$. Now, consider how much $I(\alpha_0|\alpha_0)$ exceeds the information attached to some other parameter, β, value:

$$I(\alpha_o|\alpha_o) - I(\beta|\alpha_o) = \int_y \log\left\{f\left(y; \alpha_0\right) / f(y; \beta)\right\} f(y; \alpha_0) dy \tag{2.31}$$

This equation is known as Kullback-Leibler information gain [55]. Since log is a concave function, Jensen's inequality implies that

$$-\int_y \log\{f(y;\beta)/f(y;\alpha_0)\}f(y;\alpha_0)dy \leq -\log\int_y \{f(y;\beta)/f(y;\alpha_0)f(y;\alpha_0)dy$$
$$= -\log\int_y f(y;\beta)dy = 0.$$

Therefore $I(\alpha_o|\alpha_o) - I(\beta|\alpha_o)$ is always non-negative and $I(\beta|\alpha_0)$ as a function of β attains its maximum at the true value $\beta = \alpha_0$. The entity $I(\alpha_o|\alpha_o) - I(\beta|\alpha_o)$ denotes the distance from $f(y;\beta)$ to $f(y;\alpha_0)$ when $f(y;\beta)$ is used to approximate $f(y;\alpha_0)$. Although it is common to refer to Kullback-Leibler information gain [55] as a distance, it is not a distance in the usual geometric sense.

Statistical models can be expressed by conditional density in the form of $f(y|x;\beta)$ for Y given the observed value x of X. If we want to test the null hypothesis $H_0 : \beta = 0$ against $H_1 : \beta = \beta_0$, the distance between the two models indexed by $\beta = 0$ and $\beta = \beta_0$ can be provided by the Kullback-Leibler information gain $I(\beta_0|\beta_0) - I(0|\beta_0)$. β_0 can be replaced by $\widehat{\beta}$, a consistent estimate of β. In exponential family models where censoring is not present, a standard estimate of information gain will be provided by n^{-1} times the usual likelihood ratio test statistic (Kent (1983) [50], (1986) [51]).

R^2 and Kullback-Leibler information gain

Kullback (1951) [55] pointed out the relationship between the Kullback-Leibler information gain [55] and the correlation coefficient. Suppose that H_0 and H_1 are null and alternative hypotheses as follows. Under H_0, X and Y have bivariate normal density with mean zero, variance σ_x and σ_y respectively and ρ_{xy} correlation, and under H_1, X and Y are independent with respective probability densities $f_1(x)$ and $f_2(y)$. Now the information gain may be written as

$$I(\alpha_o|\alpha_o) - I(\beta|\alpha_o) = \int \log\{f(x,y)/f_1(x)f_2(y)\}f(x,y)dxdy$$

where

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y(1-\rho_{xy}^2)^{1/2}}\exp[-\frac{1}{2(1-\rho_{xy}^2)}(\frac{x^2}{\sigma_x^2} - 2\rho_{xy}\frac{xy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2})]$$

and the marginal densities are

$$f_1(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\exp[-\frac{x^2}{2\sigma_x^2}]$$
$$f_2(y) = \frac{1}{\sigma_y\sqrt{2\pi}}\exp[-\frac{y^2}{2\sigma_y^2}]$$

Kullback (1951) [55] showed that

$$I(\alpha_o|\alpha_o) - I(\beta|\alpha_o) = \iint \log\{f(x,y)/f_1(x)f_2(y)\}f(x,y)dxdy = -\frac{1}{2}\log(1-\rho_{xy}^2) \quad (2.32)$$

so that $I(\alpha_o|\alpha_o) - I(\beta|\alpha_o)$ is a function of the correlation coefficient ρ_{xy} only , and ranges from 0 to ∞ as $|\rho|$ ranges from 0 to 1.

Kent (1983) [50] generalised equation (2.32) by defining a measure of correlation between two random variables. He proposed a measure of correlation for more general models as

$$\rho_{IG}^2 = 1 - \exp\{-\Gamma(H_1 : H_0)\}. \quad (2.33)$$

where $\Gamma(H_1 : H_0) = 2\{I(\alpha_0;\alpha_o) - I(\beta;\alpha_o)\}$ is twice the Kullback-Leibler information gain and $I(\alpha_0;\alpha_o)$ is the expected information assuming H_0 and $I(\beta;\alpha_o)$ is the expected information assuming H_1. Note that a factor of 2 is introduced to generalise this measure since we have factor $\frac{1}{2}$ in equation (2.32). Therefore, the correlation coefficient ρ_{IG}^2 has the following properties:

1) If X and Y are independent, then $\rho_{IG}^2 = 0$.

2) $0 \leqslant \rho_{IG}^2 < 1$.

Explained randomness measures in survival models

Measures in this category use the relationship proposed by Kent (1983) [50] in equation (2.33) to provide a measure of predictive ability in survival models. This category includes measures proposed by Maddala (1983) [67], Kent and O'Quigley (1988) [49], Magee (1990) [68], Nagelkerke (1991) [71], Verweij and Van Houwelingen (1993) [113], Xu and O'Quigley (1999) [116], and O'Quigley et al (2005) [80]. Most of the measures in this category have been introduced for the Cox PH model, and the main difference between them is in the way they construct the Kullback-Leibler information gain [55].

Maddala (1983), Magee (1990), and Nagelkerke (1991) measures - ρ_n^2

Maddala (1983) [67] and Magee (1990) [68] proposed two similar measures for the models that use maximum likelihood as a criterion of fit. Maddala (1983) proposed

$$\rho_M^2 = 1 - \{L(\beta)/L(0)\}^{2/n}$$

where $L(\beta)$ and $L(0)$ denote the likelihoods of the fitted and the null models, respectively. Magee's (1988) proposed measure is

$$\rho_n^2 = 1 - \exp\left\{-\frac{2}{n}(l_\beta - l_0)\right\}$$

where l_β and l_0 are the log likelihoods of the fitted and the null models, respectively. Since Nagelkerke (1991) [71] studied the properties of these measures, they are generally attributed to him in the literature. Allison (1995) [3] suggested ρ_n^2 for survival models, including the Cox PH model.

Kent (1986) [51] showed that for the exponential family, an estimate of information gain is provided by n^{-1} times the usual likelihood ratio test statistic. Therefore, ρ_n^2 is a measure of randomness for exponential family models. It is unclear what ρ_n^2 measures in the context of survival models and in particular the Cox PH model. ρ_n^2 can be written as

$$\rho_n^2 = \frac{\exp\left(-\frac{2}{n}l_0\right) - \exp\left(-\frac{2}{n}l_\beta\right)}{\exp\left(-\frac{2}{n}l_0\right)}$$

where $\exp\left(-\frac{2}{n}l_0\right)$ and $\exp\left(-\frac{2}{n}l_\beta\right)$ are defined as the randomness of outcome in the null model and the randomness of outcome given the covariate. Then ρ_n^2 can be interpreted as the proportion of randomness in the outcome which is explained by the covariate.

Verweij and Houwelingen (1993) [113] proposed a similar measure to ρ_n^2 in which the log likelihoods, l_β and l_0, are replaced with the cross-validated log likelihood counterparts, i.e. cvl_β and cvl_0.

$$\rho_{cv}^2 = 1 - \exp\left(-\frac{2}{n}(cvl_\beta - cvl_0)\right).$$

Xu and O'Quigley measure (2005)

Xu and O'Quigley (2005) [80] suggested a modified measure of explained randomness for the Cox PH model. They proposed replacing sample size, n, in ρ_n^2 by the number of

events, i.e. effective sample size, k:

$$\rho_k^2 = 1 - \exp\left\{-\frac{2}{k}(l_\beta - l_0)\right\}.$$

Kent and O'Quigley measures (1988) - ρ_W^2 & $\rho_{W.A}^2$

Kent and O'Quigley (1988) [49] proposed a different measure of explained randomness for the Cox PH model. The Cox PH model in (2.15) can be written as

$$f(t|X;\beta) = h_0(t)\exp\left\{\beta X - e^{\beta X}\int_0^t h_0(u)du\right\}. \tag{2.34}$$

The underlying distribution remains unknown in the Cox PH model, i.e. the baseline hazard $h_0(t)$ is completely unspecified, which makes the construction of Kullback-Leibler information gain [55] impossible. The measures proposed by Kent and O'Quigley (1988) [49], Xu and O'Quigley (1999) [116], and O'Quigley et al (2005) [80] make use of properties of the Cox PH model, as explained below, to find a way round this problem.

Kent and O'Quigley's (1988) measure replaces the baseline hazard, $h_0(t)$, in the Cox PH model with a monotonic function of time to form Kullback-Leibler information gain [55]. Any transformation of time which does not change the rank of event or censoring times will result in the same parameter estimates in the Cox PH model. Kent and O'Quigley (1988) [49] claimed that since this does not change the parameter estimates in the model, it should not change the predictive ability of the model either. Thus, any strictly monotonic transformation of time, $T^* = \phi(T)$, gives the same regression coefficient in the Cox PH model as T does. Kent and O'Quigley utilized this property of the Cox model and defined $h_0^*(t) = \alpha\exp(\mu)t^{\alpha-1}$ for any choice of μ and α. By choosing this baseline hazard we ensure that the baseline hazard is proportional to a power of t.

Kent and O'Quigley (1988) [49] argued that generally $h_0(t)$ can be replaced by any other strictly monotonic transformation of t, but the expected log likelihood function, $I(\beta;\beta)$ and $I(0;\beta)$, calculations will be too awkward to compute easily. Note that finding a suitable transformation would in practice not be possible if no parametric form for baseline hazard was assumed. Therefore, if we replace $h_0(t)$ in (2.34) with $h_0^*(t)$, the conditional distribution of T^* given $X = x$, $f^*(t|X;\beta)$, follows a Weibull distribution:

$$f^*(t|X;\beta) = \alpha\exp(\mu+\beta X)t^{\alpha-1}\exp\left[-t^\alpha\exp\left(\mu+\beta X\right)\right].$$

Now it is possible to construct expected log likelihoods under the full and null models, $I(\beta;\beta)$ and $I(0;\beta)$, and hence the Kullback-Leibler information gain [55], $\Gamma = 2\{I(\beta;\beta) - I(0;\beta)\}$. Time, t, is indirectly involved in the calculation of $I(\beta;\beta)$ and $I(0;\beta)$ and is used only through the estimate of β, i.e. $\widehat{\beta}$. $I(\beta;\beta)$ has a closed form but $I(0;\beta)$ should be numerically maximised to evaluate the Kullback-Leibler information gain [55]. Using equation (2.33) developed by Kent (1983), Kent and O'Quigley (1988) proposed the following measure of explained randomness/uncertainty for the Cox PH model:

$$\rho^2_W = 1 - \exp(-\Gamma). \tag{2.35}$$

Since no explicit formula is available for ρ^2_W, they proposed an approximation,

$$\rho^2_{W,A} = \frac{\mathbf{Var}_X(\beta' x)}{\mathbf{Var}_X(\beta' x) + 1} \tag{2.36}$$

which is numerically easier to compute.

Note that replacing the baseline hazard function with $h^*_0(t) = \alpha \exp(\mu) t^{\alpha-1}$ in the Cox PH model means that the conditional distribution of T given X follows a Weibull distribution. Therefore $Y = \ln T^*$ follows a linear regression model:

$$Y = \ln(T^*) = -\sigma(\mu + \beta X) + \sigma\varepsilon$$

where $\sigma = \alpha^{-1}$, ε is independent of X, Y has density $f(y)$ where

$$f(y) = e^y \exp(-e^y),$$

i.e. the extreme value (Gumbel) density (Lawless, 1982 [59]) with variance $\frac{\pi^2}{6} \simeq 1.645$. In the extreme value (Gumbel) density, σ and μ are scale and location parameters, respectively. Therefore a measure of explained variation for the Cox PH model, based on Helland's measure (1987) in equation (2.23) is:

$$R^2_{PM} = \frac{\mathbf{Var}_X(\beta' x)}{\mathbf{Var}_X(\beta' x) + 1.645}. \tag{2.37}$$

Kent and O'Quigley (1988) [49] suggested that if $\mathbf{Var}_X(\beta' x)$ is small, there exists the following relationship between explained variation measure, R^2_{PM}, and explained ran-

domness measure, $\rho^2_{W,A}$, in the Cox PH model.

$$\rho^2_{W,A} = 1.645 R^2_{PM} \tag{2.38}$$

Xu and O'Quigley measure (1999) - ρ^2_{XuOQ}

Xu and O'Quigley (1999) [116] proposed an alternative measure to Kent and O'Quigley's measure (1988) [49] based on the Kullback-Leibler information gain [55]. They argued that this alternative measure is more natural in the context of the Cox proportional hazards regression. An apparently unusual feature of the O'Quigley and Flandre (1994) [75] measure of explained variation in section 2.3.1 is that, rather than measuring the ability of the covariate to predict time, as we might expect, we measure how close model-based covariate predictions are to the observed value of covariate at each failure time. O'Quigley and Xu (1999) [116] showed that doing things this way around is, in fact, natural in the context of the Cox PH model and amounts to predicting not times themselves, but, instead, the time ranks or the ordering of the observations.

Xu and O'Quigley (1999) [116] used the above property and defined an alternative Kullback-Leibler information gain [55] to propose a new measure of explained randomness. Recall the Cox PH model

$$f(t|X;\beta) = h_0(t) \exp\left\{\beta X - e^{\beta X} \int_0^t h_0(u)du\right\}.$$

Xu and O'Quigley (1999) [116] indicated that we could equally work with an alternative $I_2(\beta;\beta)$ given by

$$I_2(\beta;\beta) = \int_T \int_X \log\left\{g(x|t;\beta)\right\} g(x|t;\beta) dx dF(t) \tag{2.39}$$

where $F(t)$ is the marginal distribution function of T, and $g(x|t;.)$ is the conditional density or conditional probability function of X given T. As before, define the Kullback-Leibler information gain [55] as $\Gamma_2(\beta) = 2\{I_2(\beta;\beta) - I_2(0;\beta)\}$ and

$$\rho^2_{XuOQ}(\beta) = 1 - \exp(-\Gamma_2(\beta)). \tag{2.40}$$

Xu and O'Quigley (1999) [116] showed that under the proportional hazards model

the conditional distribution of X given T, $g(x|t;\beta)$, is consistently estimated by

$$\widehat{P}(X \leqslant x|T=t) = \sum\nolimits_{\{i:X_i \leqslant x\}} \pi_i(t;\widehat{\beta})$$

where

$$\pi_i(t;\beta) = \frac{Y_i(t)\exp(\beta X_i)}{\sum_{l=1}^{n} Y_l(t)\exp(\beta X_l)} \tag{2.41}$$

is the conditional probability of choosing individual i, given all the individuals at risk at time t and that one individual is to be selected to fail. $Y_i(t)$ in $\pi_i(t;\beta)$ is at risk indicator for individual i. The product of the π_is over the observed failure times is the partial likelihood (Cox (1972) [19] and Cox (1975) [18]).

Let $t_1 < ... < t_k$ be distinct failure times. We estimate the conditional distribution of X given T by $\left\{\pi_j(t;\widehat{\beta})\right\}_j$, $j = 1, ..., k$, and the marginal distribution of T, $F(t)$, by the Kaplan-Meier estimate. Let $W(t_j)$ be the jump of the Kaplan-Meier curve at time t_j. Then

$$\Gamma_2(\beta) = 2\int_T \int_X \log\left\{\frac{g(x|t;\beta)}{g(x|t;0)}\right\} g(x|t;\beta)dxdF(t)$$

can be consistently estimated by

$$\widehat{\Gamma}_2(\widehat{\beta}) = 2\sum\nolimits_{j=1}^{k} W(t_j)\sum\nolimits_{i=1}^{n} \pi_i(t_i;\widehat{\beta})\log\left\{\frac{\pi_i(t_i;\widehat{\beta})}{\pi_i(t;0))}\right\}$$

where the outer sum is effectively over those subjects that are in the risk set at time t_j. The above expression can be shown to be equal to

$$\widehat{\Gamma}_2(\widehat{\beta}) = 2\sum\nolimits_{j=1}^{k} W(t_j)\left\{\widehat{\beta}\frac{\sum_{l=1}^{n} Y_l(t)X_l\exp(\widehat{\beta}X_l)}{\sum_{l=1}^{n} Y_l(t)\exp(\widehat{\beta}X_l)} - \log\frac{\sum_{l=1}^{n} Y_l(t)\exp(\widehat{\beta}X_l)}{\sum_{l=1}^{n} Y_l(t)}\right\}.$$

Then using Kent's (1983) formula, we have:

$$\widehat{\rho}^2_{XuOQ} = 1 - \exp\left\{-\widehat{\Gamma}_2(\widehat{\beta})\right\}.$$

The results of Xu and O'Quigley's investigation [116] indicate that this measure is an approximation to the Kent and O'Quigley measure (1988) [49] and O'Quigley and Flandre measure (1994) [75].

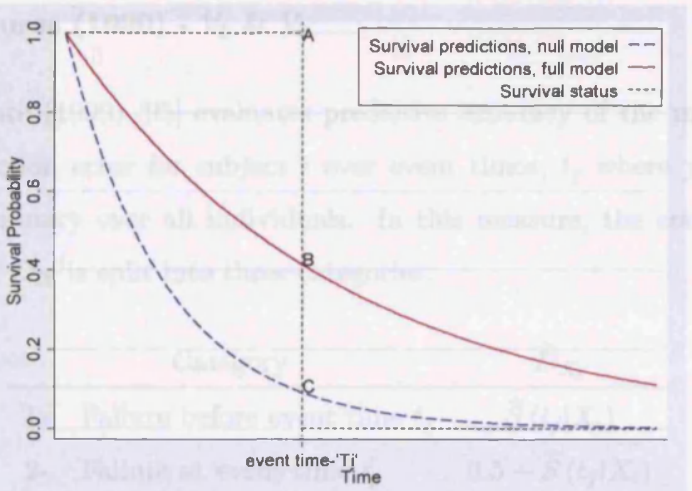


Figure 2-3: Schematic presentation of survival status (dotted line), survival predictions from the null model (broken line), survival prediction given covariates (solid line) for individual i in predictive accuracy measures.

2.3.3 Measures of predictive accuracy

The third category, predictive accuracy measures, includes three main measures proposed by Schemper (1990) [95], Graf et al (1999) [31], and Schemper and Henderson (2000) [97]. Although they use different mechanisms to evaluate the predictive ability of a survival model, they broadly quantify the accuracy of predicted survival probabilities. In normal linear regression R^2 quantifies how close the model-based predictions are to the observed values of the outcome. The proposed measures in this category apply this to survival models, but on the survival probability scale rather than the actual time scale. Figure 2-3 shows schematically the mechanism that these measures apply to quantify the predictive accuracy. The dotted line is the survival status of an individual who survived until event time t_j; therefore its survival status is 1 until t_j and 0 thereafter.

In these measures, predicted survival probabilities from the null and full models are compared with the survival status of individuals, i.e. 1 if alive at time t and 0 otherwise, at each time point t_j in the observation period which result in the marginal and conditional prediction errors, i.e. AB is compared to the AC in figure 2-3. This leads to a measure which quantifies relative gain in terms of predictions when using covariates. The main difference between these measures is in the specification of the distance function, D, that they use to penalise the marginal and conditional prediction errors.

Schemper measures (1990) - V_1 & V_2

Schemper's measure (1990) [95] evaluates predictive accuracy of the model by first averaging the prediction error for subject i over event times, t_j where $j = 1, ..., k$, then averaging this summary over all individuals. In this measure, the contribution to the distance functions D_X is split into three categories:

	Category	$\widehat{D}_{X_i}$
1-	Failure before event time t_j	$\widehat{S}(t_j \mid X_i)$
2-	Failure at event time t_j	$0.5 - \widehat{S}(t_j \mid X_i)$
3-	Failure after event time t_j	$1 - \widehat{S}(t_j \mid X_i)$

Where $\widehat{S}(t_j|X_i)$ is the estimated survival probability at time point t_j for individual i given covariate X_i. D is defined in a similar fashion using marginal survival probabilities. Schemper's measure (1990), V_1, is calculated using the following formula.

$$\widehat{V}_1 = \frac{\sum_{i=1}^{n}(\frac{1}{k_i}\sum_{failures\ t_j}\widehat{D}) - \sum_{i=1}^{n}(\frac{1}{k_i}\sum_{failures\ t_j}\widehat{D}_{X_i})}{\sum_{i=1}^{n}(\frac{1}{k_i}\sum_{failures\ t_j}\widehat{D})}$$

Schemper (1994) [96] also proposed V_2 measure similar to V_1 which is defined in terms of squared sums $(\frac{1}{k_i}\sum_{failures\ t_j}\widehat{D})^2$ and $(\frac{1}{k_i}\sum_{failures\ t_j}\widehat{D}_{X_i})^2$. O'Quigley et al (1999) [77] studied the population characteristics of Schemper measures mathematically.

Graf et al (1999) $R^2_G(T^*)$ & Schemper and Henderson (2000) V_{SchH} measures

In contrast to Schemper measures (1990) V_1 and V_2, Graf et al measure (1999) [31] and Schemper and Henderson measure (2000) [97] calculate the predictive accuracy of the model by first calculating the average prediction error at event time t_j, $j = 1, ..., k$ for all individuals, then taking a weighted average of this summary over the event times in the observation period. These two measures use a weighting scheme to compensate for the loss of information due to censoring.

In Graf's measure (1999), first a particular time point, T^*, at which we would like to assess the survival probability predictions is specified. The time point should be equal or before the last failure time in the data. In this measure, the contributions to the distance function, $D_x(T^*)$, are split into three categories:

	Category	$\widehat{D}_{X_i}(T^*)$
1-	Failure before T^*	$\widehat{S}(T^*\|X_i)/\widehat{G}(T^*)$
2-	Censored before T^*	0
3-	Failure or censored after T^*	$(1-\widehat{S}(T^*\|X_i))/\widehat{G}(T^*)$

where $\widehat{S}(T^*|X_i)$ is the estimated survival probability at time point T^* for individual i given covariate X_i, and $\widehat{G}(T^*)$ is the estimated survivor function for censoring times. Then the average squared distance is calculated

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{D}^2_{X_i}(T^*). \tag{2.42}$$

Compare this with its counterpart calculated from the model without covariate which leads to Graf at al measure (1999)

$$\widehat{R}^2_G(T^*) = \frac{\sum_{i=1}^{n}\widehat{D}^2_i(T^*) - \sum_{i=1}^{n}\widehat{D}^2_{X_i}(T^*)}{\sum_{i=1}^{n}\widehat{D}^2_i(T^*)}.$$

Schemper and Henderson measure (2000) [97] is based on a similar principle to Graf et al's measure (1999) [31]. The main difference between them is that in the second category where Schemper and Henderson measure (2000) [97] uses the proposed regression model to determine the probability of reaching T^* if censored earlier. Therefore, the contribution to distance function will be as follows

	Category	$\widehat{D}_{X_i}(T^*)$
1-	Failure before T^*	$\widehat{S}(T^*\|X_i)$
2-	Censored before T^*	$P*(1-\widehat{S}(T^*\|X_i))+(1-P)*\widehat{S}(T^*\|X_i)$
3-	Failure or censored after T^*	$(1-\widehat{S}(T^*\|X_i))$

where $\widehat{P} = \widehat{S}(T^*|X_i)/\widehat{S}(T|X_i)$ is the probability of reaching T^* if censored earlier. Unlike Graf et al's measure (1999) [31] which evaluates the predictive accuracy at a time point in the study period, i.e. T^*, this measure gives a summary measure over all failure times. In addition, Graf at al's measure (1999) averages squared distance over all individuals, but Schemper and Henderson's measure (2000) averages absolute distance, $\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_{X_i}(T^*)$ over all individuals at an event time.

To give a summary measure, a weighted average is taken over all failure times

$$\widehat{D}_{SH}(X) = \frac{\sum_{failures\ t_j} \left\{ \frac{1}{n} \sum_{i=1}^{n} \widehat{D}_{X_i}(t_j) / \widehat{G}(t_j) \right\}}{\sum_{failures\ t_j} \left(1/\widehat{G}(t_j) \right)}. \tag{2.43}$$

Compare this with its counterpart calculated from the model without covariate to give an estimate of the predictive accuracy measure.

$$\widehat{V}_{SchH} = \frac{\widehat{D}_{SH} - \widehat{D}_{SH}(X)}{\widehat{D}_{SH}}. \tag{2.44}$$

2.3.4 Other proposed measures in survival models

Two measures proposed by Harrell (1986) [35] and Schemper and Kaider (1997) [98] form the other category. They do not belong to any of the three main categories discussed above. Harrell's measure (1986) [35] is the proportion of increase in the log likelihood of the model given covariates compared with the one from the null model, given by:

$$R^2_{HL} = \frac{l_0 - l_\beta}{l_0}$$

where l_0 and l_β are the log likelihoods under the null and full model, i.e. model with covariates, respectively.

Schemper and Kaider (1997) [98] apply multiple imputation methods to impute the censored survival times and make use of a nonparametric measure of correlation, such as Spearman correlation coefficient (Spearman (1904) [108]) or Kendall τ (Kendall (1938) [47]), to calculate a measure of association between completed or imputed survival times and covariates in the Cox PH model. The algorithm that is applied to compute this measure is included in Appendix B. We identify this measure as R^2_{SchK} throughout this thesis.

2.4 Discussion

Over the course of this chapter we have presented the proposed measures of predictive ability for survival models, in particular the measures proposed in the context of the Cox PH model. In the linear regression, R^2 is a well known measure of predictive ability. Different interpretations of R^2 as a measure of explained variation, explained randomness,

and predictive accuracy has led to a wide range of measures for survival models.

The diversity of the measures proposed for survival models raises two important questions: I) what does each of these measures estimate; II) which measures are recommendable for general use in practical applications. To address these questions, we need to compare the proposed measures systematically to have a better understanding of their performance. This requires a set of criteria against which we can evaluate the proposed measures. The criteria should indicate the properties that a measure of predictive ability should possess in the context of survival analysis. These criteria will provide us with a framework for comparing the proposed measures consistently and help us to investigate the behaviour of the measures in similar conditions.

The next chapter presents the criteria that, in our opinion, a measure of predictive ability should possess in the context of survival analysis. Then, we compare the proposed measures with regard to these criteria. This will suggest the need for further empirical studies to investigate the performance of the proposed measures.

Chapter 3

Investigation of the proposed measures

3.1 Introduction

As reviewed in the previous chapter, many measures have been proposed to assess the predictive ability of survival models. An extensive literature search uncovered only one study to evaluate the performance of selected R^2 analogues in survival models. Schemper and Stare (1996) [99] reviewed some of the predictive ability measures in survival analysis with the conclusion that no particular statistic can be recommended for general use, mainly due to the impact of censoring on the measures. They described several properties that a measure should have, such as independence from censoring, intuitive interpretation, and robustness against model mis-specification. Nagelkerke (1991) also presented some properties for a measure in more general models such as consistency with classical R^2 in linear regression, intuitive interpretation, dimensionless, i.e. it does not depend on the units, and independence of sample size.

In this chapter, we define a set of criteria that a suitable measure of predictive ability should possess in the context of survival analysis. These criteria are used as a basis to compare the proposed measures. The criteria provide us a framework for choosing a suitable measure of predictive ability. Some of the proposed properties have been recommended in previous work by Schemper and Stare (1996) [99] and Royston and Sauerbrei (2004) [93].

3.2 Properties of a "good" measure

The properties that a suitable measure of predictive ability should possess can be classified into two main categories: essential properties; and desirable properties. Essential properties are the ones that a suitable measure of predictive ability should possess in the context of survival analysis. Satisfying desirable properties could give a measure an advantage over the other measures.

3.2.1 Essential properties

I) Independence from censoring:

The expected value of the measure should be approximately independent of the amount of censoring. Censoring is one of the basic properties of survival data and it is present in almost all practical applications. Therefore, a measure that is unduly affected by the amount of censoring is considered unsuitable.

Since one of the aims of this study is to make practical comparisons of the measures, we quantify the extent of censoring effect by comparing the average percentage change in the expected value of the measures compared to that of non-censored data. We translate the resulted figures into 4 categories, each representing the extent of censoring: 1) censoring has almost no effect; 2) censoring has a slight effect; 3) censoring has moderate effect; 4) censoring has large effect. For further definition of these categories, see section 5.3.

II) Independence from sample size:

Sample size should not affect the measure. The measure should converge to the population/true value of the measure, if they exist, in both censored and non-censored data.

III) Monotonicity:

a) Parameter monotonicity: An appropriate measure of predictive ability should acquire higher values as the effect of prognostic factor, i.e. covariate, on the outcome variable becomes stronger. This means that the expected value of the measure does not decrease as the absolute value of the parameter estimate in the model increases.

b) Number of variables monotonicity: An appropriate measure of predictive ability should acquire higher values as new prognostic factors, i.e. covariates, are included in the model. The number of variables monotonicity means that the expected value of a suitable measure of predictive ability should not decrease by adding new covariates.

3.2.2 Desirable properties

A measure of predictive ability which possesses desirable properties could have an advantage over the other measures.

I) Robustness:

In normal linear regression, R^2 is influenced by the outliers and extreme observations (Kvalseth (1985) [56], Montgomery (2001) [69]). The impact of such observations has not been studied in the proposed measures of predictive ability for survival models. It is worth bearing in mind that unlike linear regression, where there are established methods to identify such observations in the model building process and rectify them, no method has been universally accepted in survival analysis to identify outliers and extreme observations. Therefore, a measure that is resistant to such observations might have an advantage over the other measures.

II) Confidence intervals:

Confidence intervals show how much uncertainty is associated with point estimates. A measure is preferred if its confidence intervals can be obtained, although using the bootstrap is always a possibility.

III) Partial R^2:

Partial R^2 measures the correlation between outcome variable, for example survival time, and a covariate when other covariates in the model are held constant with respect to the outcome variable and that covariate. It measures the marginal contribution of one covariate when all the others are already included in the model. It can help us to examine the relative importance of different sets of covariates.

IV) Adjusted R^2 :

It is also desirable to provide an adjusted R^2 for the reasons explained in section 2.1.3.

V) Generalisability:

Ideally, the measure should be generalisable for different types of survival models. Parametric survival models sometimes are considered as alternatives to the Cox PH model, and they might give better fit to the data under study. Therefore, a flexible measure can be used in these circumstances. Some other areas of extension include flexible parametric models proposed by Royston and Parmar (2002) [92], as presented in section 2.2.2.

3.3 Shortcomings of some measures

An extensive literature search uncovered properties of the proposed measures with regard to the criteria outlined in the last section. One of the objectives in this thesis is to recommend one or more measures for general use. When comparing the proposed measures with regard to the essential properties, the evidence from previous studies suggests that some of the proposed measures are unsuitable for survival models. These measure generally do not satisfy the essential properties.

We classify the proposed measures of predictive ability into two categories: potentially recommendable measures; and unsuitable measures. Potentially recommendable measures are those for which the previous studies do not provide evidence against the essential criteria outlined in section 3.2, or the evidence is inconclusive to reach a definite conclusion when comparing the measures against these properties.

The class of unsuitable measures includes Korn and Simon (1990) [53], Schemper's V_1 and V_2 (1990) (1994) [95] [96], Akazawa (1997) [2], Harrell's likelihood (1984), Maddala (1983) [67], Magee (1990) [68], Nagelkerke (1991) [71] and Verweij & Van Houwelingen (1993) [113].

Despite some promising properties presented by Korn and Simon (1990) (1991) [53] [54] and Henderson (1995) [42], Korn and Simon's class of measures (1990) [53] is classified as unsuitable because the previous studies suggest that the amount of censoring has a considerable impact on the measures. Schemper and Stare (1996) [99] performed a range of simulation studies on the measures, as well as some others. They gener-

ated exponentially distributed survival times ($f(t) = \lambda \exp(-\lambda t)$) with hazard function $\lambda = exp(-X\beta)$, respectively with β set to 0, log(2), log(16) and log(64). For example, the results of simulation for β=log(16) showed the expected values of the measure as 0.35, 0.56 and 0.19 for 0%, 50% and 90% censored data, respectively. In addition, in a simulation study Stare (1994) ([109], page 33) showed that the measure is not invariant under monotonic transformation of time. Therefore, whilst a monotonic transformation of time will not change the parameter estimates in the Cox PH model, Korn and Simon's measure will result in a different value in the Cox PH model.

Schemper's V_1 and V_2 measures (1990) (1994) [95] [96] are influenced heavily by the degree of censoring. Schemper and Stare (1996) [99] investigated the effect of censoring on V_1. The expected value of the measure, resulted from the simulation study with the same setting as the Korn and Simon's measure explained above, were 0.59, 0.46 and 0.11, respectively.

Akazawa's measure (1997) [2] is also classified as unsuitable for two reasons. First, this measure is based on the rather strong assumption that the follow-up terminates at some prespecified time with "no loss to follow-up". Censoring is one of the basic properties of survival data and it is present in almost all practical applications. Accepting the assumption of "no loss to follow-up" makes this measure inapplicable in practice for a vast majority of studies. Second, this measure is heavily influenced by the degree of censoring. Akazawa (1997) [2] performed a simulation study with exponentially distributed survival times and one dichotomous covariate for a range of hazard ratios. The results were presented in graphs ([2], pages 233 & 234) which suggest that this measure is heavily influenced by the amount of censoring.

Harrell 's measure (1986) [37] which is based on likelihood function is another unsuitable measure. This measure is slightly influenced by the degree of censoring and heavily influenced by the changes in sample size. Stare (1994) [109] performed a simulation study on exponentially generated survival data with one discrete covariate. The value of the measure for non-censored data was 0.07 for a hazard ratio of 16 which increased to 0.11 for 50% censored data. The expected value of the measure decreased as the sample size increased; the expected value of the measure were 0.11, 0.07 and 0.06 for sample sizes 80, 800, 4000, respectively.

Nagelkerke's measure (1991) [71] is also considered unsuitable for survival models. O'Quigley et al (2005) [80] generated a large sample ($n = 5000$) of exponentially dis-

tributed survival times ($f(t) = \lambda \exp(-\lambda t)$) with hazard function $\lambda = \exp(-X\beta)$, respectively with β set to 0, log(2), log(16) and log(64). The results of simulation for β=log(16) showed the expected value of the measure as 0.68, 0.64, and 0.13 for 0%, 50%, and 90% censored data, respectively. The measures proposed by Maddala (1983) [67] and Magee (1990) [68], therefore, are heavily influenced by the degree of censoring. Verweij and Van Houwelingen's measure (1993) [113] will have the same drawback since it is very similar to Nagelkerke's measure (1991) [71] with one difference; it uses cross-validated log-likelihoods instead of log-likelihoods.

3.4 Tables of the properties of measures

Tables 3.1 to 3.5 present potentially recommendable measures and unsuitable measures, their properties, and the programs available to compute them. The description of the properties are at the bottom of each table. Tables 3.1 and 3.2 present potentially recommendable measures and their status regarding the essential and desirable properties. Unsuitable measures have been rejected on the grounds that they did not satisfy all the essential properties outlined in section 3.2.

Key of the tables

The description of the terms used in the tables are as follows:

- nk: not known - the evidence from previous studies is inconclusive.
- yes: the measure does possess the desired property.
- no: the measure does not possess the desired property.

Table 3.1: Summary of the essential properties of the potentially recommendable measures of predictive ability in survival analysis

Measure category	Measures	I	II	III	
				a	b
Explained variation	Helland;Kent & O'Quigley,R^2_{PM}	nk	nk	nk	nk
	O'Quigley & Flandre (94)	nk	nk	yes	yes
	Xu and O'Quigley (01)	yes	nk	yes	yes
	Royston & Sauerbrei (04)	nk	nk	yes	nk
	Royston (06)	nk	nk	yes	yes
Explained randomness	Kent & O'Quigley (88)	yes	nk	yes	yes
	Xu & O'Quigley (99)	yes	nk	yes	yes
	O'Quigley et al (05)	nk	nk	yes	yes
Predictive Accuracy	Graf et al (99)	nk	nk	nk	nk
	Schemper & Henderson (00)	nk	yes	yes	nk
Other	Schemper & Kaider (97)	yes	yes	nk	nk

I) Independence from censoring: the expected value of the measure should be approximately independent of the degree of censoring.

II) Independence from sample size: sample size should not affect the measure.

III) Monotonicity

a) Parameter monotonicity: the measure should not decrease as the absolute value of the parameter estimates increase.

b) Number of variables monotonicity: the measure does not go down by adding new covariates to the model.

Table 3.2: Summary of the desirable properties of the potentially recommendable measures of predictive ability in survival analysis

Category	Measure	I	II	III	IV	V
Explained variation	Helland;Kent & O'Quigley,R^2_{PM}	nk	yes	yes	yes	yes
	O'Quigley & Flandre (1994)	nk	yes	nk	yes	nk
	Xu and O'Quigley (01)	nk	yes	nk	yes	nk
	Royston & Sauerbrei (2004)	yes	yes	yes	yes	yes
	Royston (2006)	nk	nk	yes	yes	yes
Explained randomness	Kent & O'Quigley (1988)	nk	yes	nk	yes	nk
	Xu & O'Quigley (1999)	nk	yes	nk	yes	nk
	O'Quigley et al (2005)	nk	nk	nk	yes	yes
Predictive Accuracy	Graf et al (1999)	nk	nk	nk	nk	yes
	Schemper & Henderson (2000)	nk	nk	nk	nk	yes
Other	Schemper & Kaider (97)	nk	nk	nk	nk	yes

I) Robustness: the measure should not be unduly affected by outliers and extreme observations.

II) Confidence intervals

III) Adjusted R^2

IV) Partial R^2

V) Generalisability: the measure should be generalisable for different types of survival models.

Table 3.3: Summary of the essential properties of the unsuitable measures of predictive ability in survival analysis

Category	Measure	I	II	III	
				a	b
Explained variation	Korn & Simon (1990)	no	nk	nk	nk
	Akazawa (1997)	no	nk	yes	nk
Explained randomness	Nagelkerke (1991)	no	nk	yes	yes
	Magee & Maddala (1990)	no	nk	yes	yes
	Verweij & Houwelingen (1993)	no	nk	nk	nk
Predictive accuracy	Schemper (1990)	no	no	nk	nk
Other	Harrell (1984)	no	no	yes	yes

I) Independence from censoring: the expected value of the measure should be approximately independent of the degree of censoring.

II) Independence from sample size: sample size should not affect the measure.

III) Monotonicity

a) Parameter monotonicity: the measure should not decrease as the absolute value of the parameter estimates increase.

b) Number of variable monotonicity: the measure does not go down by adding new covariates.

Table 3.4: Summary of the desirable properties of the unsuitable measures of predictive ability in survival analysis

Category	Measure	I	II	III	IV	V
Explained variation	Korn & Simon (1990)	nk	no	nk	nk	yes
	Akazawa (1997)	nk	no	nk	nk	nk
Explained randomness	Nagelkerke (1991)	nk	nk	nk	yes	yes
	Magee & Maddala (1990)	nk	nk	nk	yes	yes
	Verweij & Houwelingen (1993)	nk	nk	nk	yes	yes
Predictive accuracy	Schemper (1990)	nk	no	nk	nk	yes
other	Harrell (1984)	nk	no	nk	yes	yes

I) Robustness: the measure should not be unduly affected by outliers and extreme observations.

II) Confidence intervals

III) Adjusted R^2

IV) Partial R^2

V) Generalisability: the measure should be generalisable for different types of survival models.

3.5 Discussion

In this chapter we first set out the criteria that a suitable measure of predictive ability should possess in the context of survival analysis. A suitable measure should not unduly be affected by the changes in the amount of censoring and sample size. It should also be generalisable and monotonic (as explained in section 3.2).

Overall, our investigation of proposed measures of predictive ability led us to a short-list of measures, i.e. potentially recommendable measures presented in tables 3.1 and 3.2. Evidence from previous studies helped us to reject some of the measures as potential candidates, mainly on the grounds that censoring has a considerable impact on them. As it is clear from the tables of the essential and desirable properties, the performance of the potentially recommendable measures with respect to some of the properties is still unknown, which requires further investigation. This constitutes the next stage of this thesis. For example, the evidence from previous empirical studies has been mainly inconclusive about the extent of censoring effect, specially when comparing different measures.

Further work in this area would include a series of simulation studies using different censoring mechanisms for simulating censored time-to-event data, in order to thoroughly assess the impact of censoring and its magnitude under different censoring assumptions. The influence of sample size on the measures also remains unknown. The simulation study also needs to include different sample size conditions and covariate effects to investigate their impact on the measures.

The next chapter presents simulation design followed by the assessment of measures in each category in the following chapters. The main objective of the simulation study is to provide a thorough comparison of the potentially recommendable measures with regard to the properties set out in this chapter.

Table 3.5: Summary of the programs available to calculate the proposed measures of predictive ability in survival analysis

Category	Measure	Program
Explained variation	Korn & Simon (1990)	SAS[1]
	O'Quigley & Flandre (1994)	STATA[1], R
	Xu and O'Quigley (01)	STATA[1], R
	Akazawa (1997)	Not Available
	Royston & Sauerbrei (2004)	STATA
	Royston (2006)	STATA
Explained randomness	Nagelkerke (1991)	Any
	Magee (1983) & Maddala (1990)	Any
	Verweij & Houwelingen (1993)	GAUSS
	Kent & O'Quigley (1988)	STATA[1], SAS
	Xu & O'Quigley (1999)	STATA[1], C
	Xu & O'Quigley (2005)	Any
Predictive accuracy	Schemper (1990)	SAS
	Graf et al (1999)	R, STATA
	Schemper & Henderson (2000)	SAS & R
Other	Harrell (1984)	Any
	Schemper & Kaider (1997)	SAS & R

1: program is written by the author

Chapter 4

Further assessment of the proposed measures

4.1 Introduction

The classification of proposed measures into the potentially recommendable and unsuitable measures in chapter 3 helps us to concentrate on the measures that have, so far, not been rejected as the candidate measures of predictive ability in survival models. These measures are presented in tables 3.1 and 3.2, together with their properties. The tables showed that there are still unresolved issues with regard to the proposed measures, which require further investigation. This leads us to the next stage of this study to further investigate the properties of the potentially recommendable measures. In this chapter, we recommend simulation studies to investigate the measures against the criteria for which the performance of the measures are still unknown, such as independence of censoring and sample size and monotonicity properties, i.e. parameter and number of variables monotonicity.

4.2 Limitations of previous simulation work

Although previous simulation studies have been quite informative, there is a need for further investigation to incorporate several needed refinements to more fully scrutinise the performance of the measures. First, previous simulation studies [75] [99] [116] [80] - with the exception of Schemper and Stare (1996) [99] - have not compared all the alternative

measures against each other. In our study, we carry out a comprehensive simulation study to compare the measures of predictive ability against each other. Second, prior studies have not investigated the proposed measures in the context of multiple regression. We carry out our study in the context of multiple regression. Third, no prior simulation has examined the sampling distribution of the proposed measures under different censoring types, censoring proportions, sample sizes, and covariate effects. This is done in the current study. Fourth, previous simulations have mainly studied the impact of one kind of censoring, i.e. administrative censoring [63], on the measures. We study the impact of random censoring [12] and type I censoring at an specific time, τ, as a result of constant follow-up of τ time units for all individuals. Fifth, the impact of covariate skewness has not been addressed in the previous studies. This study investigates the impact of mild to relatively high skewed covariate distributions on the measures. Sixth, outlier and extreme observations deflate or inflate R^2 in normal linear regression (Draper and Smith (1998) [21], page 246). No prior study has investigated the impact of such observations on the measures. We will carry out simulation studies investigating the impact of such observations on the measures.

4.3 Simulation study

We propose simulation studies to get a better understanding of the performance and reliability of the potentially recommendable measures of predictive ability for survival models. Simulation study provides empirical estimation of the sampling distribution characteristics rather than on theoretical expectations of those characteristics. A simulation study offers us an alternative to theoretical investigation of measures where the theoretical approach is difficult to implement, or statistical theories simply do not exist.

Although statistical theories are efficient, the validity of statistical theory is usually based on some theoretical assumptions that might be violated in the data that we have. Therefore, we are sometimes unaware of how much we can trust the theoretical estimates and how uncertain are the estimates if some crucial assumptions of the theory have been violated. We describe the study design of the simulation studies in this chapter, and present the results of simulations in the following chapters.

4.3.1 Basic steps of the study

We will take the following steps to successfully implement the simulation studies:

- Specify aims and objectives clearly
- Design the simulation study to address the unknowns in tables 3.1 and 3.2
- Generate data
- Compute the measures for survival models
- Obtain and accumulate the measures of predictive ability from each replication in survival analysis
- Analyse the accumulated measures
- Draw conclusions based on the empirical results.

4.3.2 Aims and objectives

The aims of our simulation studies are to answer questions arising with regard to the criteria which were established in chapter 3. That is to address the unknowns in tables 3.1 and 3.2 of chapter 3. The measures are mainly defined in the context of the Cox PH regression model. Therefore, the simulation study to investigate the unknowns in table 3.1 to 3.2 will be based on the assumptions of the Cox PH model. This assumption, however, affects the generalisability of the findings. Nonetheless, we use the simulation studies to disprove the measures that are less favourable when we compare them against the criteria we set up in the previous chapter.

4.3.3 Data generation

Data generation is the main part of any simulation study. To investigate the unknowns in our simulation study, we initially use techniques that are used to generate survival times in the Cox PH regression model. Leemis (1987) [61] and Leemis et al (1990) [62] presented the formula for the general relation between the hazard function and the corresponding survival time as a tool to generate survival times. Bender et al (2005) [11] presented techniques to generate survival times following distributions compatible with proportional hazards assumption such as exponential. We follow the procedure described by Bender et al (2005) [11].

4.3.4 The effect of censoring

The impact of different degrees of censoring on the potentially recommendable measures in table 3.1 is studied in our simulation studies. The impact of censoring on these measures has either been unclear or remained unknown in previous studies. The studies on the impact of censoring is implemented by considering four levels censoring proportions as 0%, 20%, 50%, and 80% censoring in the experiments. We will consider two types of censoring: random censoring [12]; and type I or administrative censoring with no staggered entry. Random censoring is rather common in clinical studies whereas type I or administrative censoring is more common in population-based studies as well as animal studies where birth cohorts are followed up until a prespecified time point, τ. This helps us to elucidate the behaviour of the measures in different censoring situations. Mechanisms to generate censored survival time observations differ in different types of censoring. They are explained in the following section.

Generating different censoring proportions

Random non-informative right censored data with a specified proportion of censored observations is generated in a similar manner to the non-censored survival times by assuming an exponential distribution for the censoring times but without including any covariates. Determining the parameters of the censoring distribution given the censoring probability is achieved by iteration. For each simulated survival time, we generate in addition a pseudo-random exponentially distributed observation representing the time to possible censoring with an specific hazard. Different choices of hazards for censoring distribution give 0, 20, 50 or 80 percent censoring on average, respectively.

A simulated survival time is treated as censored if it is greater than the corresponding simulated time from the censoring distribution. The survival times incorporating both events and censored observations are calculated for each case by combining the non-censored survival times and the censoring times. We also consider type I or administrative censoring at an specific time, τ, as a result of constant follow-up of τ time units for all individuals.

4.3.5 The effect of sample size

Different sample size conditions are considered to investigate the behaviour of these measures in different study sizes. The chosen sample sizes range from 200, quite at the low end for use with a survival model containing highly censored survival times up to 1000, a relatively large sample size. The proposed sample size conditions are 200, 500, and 1000.

4.3.6 Monotonicity effect

Increasing parameter effects (parameter monotonicity) will also be studied in the simulation study. To do this, different sizes of data sets are generated (see section 4.3.5) with a covariate with a specific distribution, i.e. normal, lognormal, and heavily skewed distributions, and exponentially distributed survival times ($f(t) = \lambda \exp(-\lambda t)$ where $\lambda = \exp(-\beta X)$) for a range of βs. The βs considered in this study are 0.223, 0.405, 0.693, and 1.386. These values result in hazard ratios of 1.25, 1.5, 2, and 4, respectively ($HR = \exp(\beta)$).

4.3.7 Survival model

The model we consider is the univariate exponential model with distribution function

$$F(t) = 1 - \exp(-\lambda t)$$

so that

$$t = F^{-1}(u) = -((\log(1-u))/\lambda)$$

then, the survival time T of the Cox PH model can be generated using the following formula:

$$T = -((\log(1-U))/\lambda_0 \exp(\beta X)))$$

where $\lambda = \lambda_0 \exp(\beta X)$ and λ_0 is the baseline or underlying hazard. The mean and variance of exponential distribution are $1/\lambda$ and $1/\lambda^2$, and the pth quantile is $t_p = -\frac{1}{\lambda}\log(1-p)$.

4.3.8 Covariate distribution

We study the measures in the context of multiple regression where prognostic index (PI), i.e. linear predictor, in the model is usually a function of several random variables. By virtue of the central limit theorem, the prognostic index should tend to Normality as the dimension of the parameter vector β increases. However, heavily skewed covariate and prognostic index distributions are not uncommon in medical research. For example, the number of positive lymph nodes and progesterone receptor in the breast cancer study used by Royston and Sauerbrei (1999) [94] are heavily skewed, with skewness 2.8 and 4.8, respectively. Furthermore, the prognostic index of the multivariate survival model that Royston and Altman (1994) [89] developed for leg ulcer is negatively skewed.

We, therefore, carry out our simulation studies with four covariate distributions. These are normal $N(0,1)$, lognormal $LN(0,1)$, and heavily skewed covariates with positive and negative skewness of 2.8 and -2.8. Fleishman (1978) [26] proposed a method based on polynomial transformation to generate sample data with desired degrees of skewness and kurtosis from standard normal distribution. We applied this method to transform the standard normal distribution to a positively skewed distribution with skewness=2.8 - graph C in 4-1 - and a heavily skewed distribution with skewness=-2.8 - graph D in 4-1. Both distributions have mean 0 and variance 1.

4.3.9 Numbers of simulations

Burton et all (2006) [12] discussed the number of simulations required and presented formulae to obtain the optimum number of runs. This depends on the degree of accuracy that is required to achieve, the true value of the estimate of interest, and the variability of estimate of interest. Burton et al (2006) [12] presented the formula

$$Number\ of\ simulations = \left(\frac{Z_{(1-\frac{\alpha}{2})}\sigma}{\delta}\right)^2$$

to determine the number of simulations required, assuming the normality of the estimated parameter. $Z_{(1-\frac{\alpha}{2})}$ is the quantile of the standard normal distribution, σ^2 is the variance for the parameter of interest, and δ is the specified level of accuracy of the estimate of interest we are willing to accept.

For example, if the true value of predictive ability measure from fitting a univariate Cox regression is 0.067 with standard deviation of 0.066, then the number of simulations

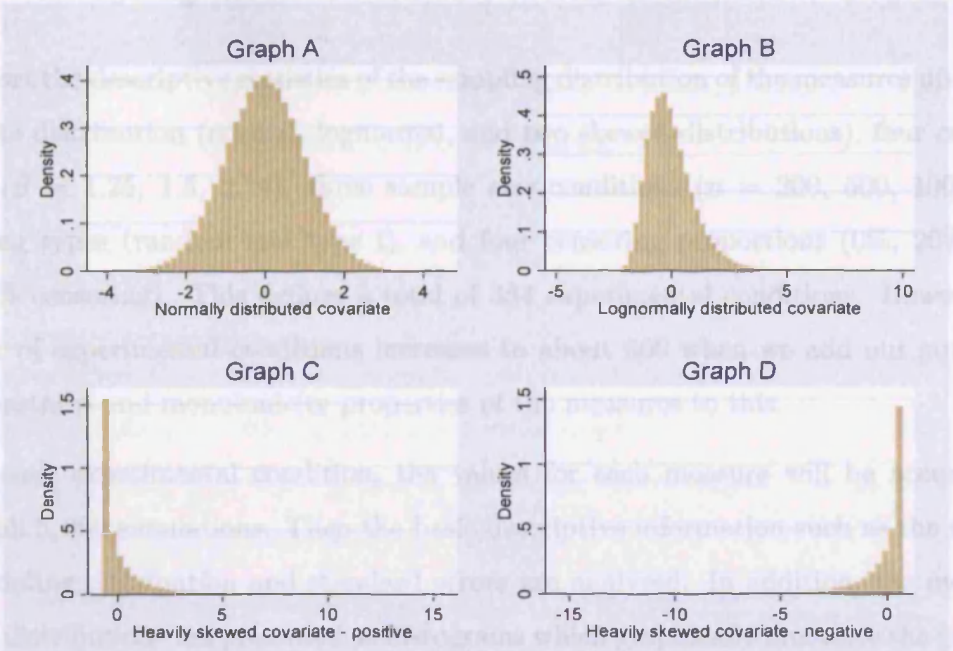


Figure 4-1: Covariate distributions considered in the simulation study

required to produce an estimate to within 5 percent accuracy of the true measure value with a 5 percent significance level would be 1492.

In this study 16 parameters are being estimated for each measure since there are four levels of covariate effect and four covariate distributions. So, the optimum number of simulations required for each parameter differs for each condition according to its variability. The number of simulations required decreases as the sample size increases because the variability of the parameter of interest decreases. We have no information about the variability of the parameters of interest in the beginning of our study to obtain the optimum number of simulations required. To be consistent, we choose 5,000 for the numbers of replications to obtain the expected value of the measures in different covariate distributions and to study the censoring and sample size effects. This might be computationally expensive, but gives us an insight into the value of the parameters and their variability. Given the results obtained in these studies, we will adjust the number of simulations later in our studies to be more efficient.

4.3.10 Analysing the accumulated statistic of interest

We report the descriptive statistics of the sampling distribution of the measures under four covariate distribution (normal, lognormal, and two skewed distributions), four covariate effects (β = 1.25, 1.5, 2, 4), three sample size conditions (n = 200, 500, 1000), two censoring types (random and type I), and four censoring proportions (0%, 20%, 50%, and 80% censoring). This defines a total of 384 experimental conditions. However, the number of experimental conditions increases to about 500 when we add our studies on the robustness and monotonicity properties of the measures to this.

At each experimental condition, the values for each measure will be accumulated across all 5,000 simulations. Then the basic descriptive information such as the mean of the sampling distribution and standard errors are analysed. In addition, the measures' sample distributions are presented as histograms which graphically illustrate the sampling distribution of the measures. This information is presented in tables for each of the measures.

4.3.11 Evaluation of the predictive ability measures

At present, very little is known about the sampling distributions of the measures in different censoring conditions. We therefore examined two key components of the performance of the measures in each experimental condition: a) mean, b) dispersion. To compare the performance of the proposed measures with regard to the amount of censoring, we calculated the difference between the expected value of the measures in different censoring conditions from the corresponding non-censored value.

As previously mentioned, since the proposed measures assess different population quantities, results to show the censoring effect on the proposed measures are computed in the relative form in each experimental condition, as a percentage of the expected value of the measures in the corresponding non-censored condition. Then we take the average across the experimental conditions. To assess the spread of the measures, we calculate the standard deviation of the sampling distribution and the coefficient of variation in each experimental condition and take the average over the experimental conditions.

4.4 Software used

Windows-based Stata and SAS are used to implement the simulations. Then, Stata is used to summarise the results and prepare the graphs.

4.5 Discussion

In this chapter we first presented the shortcomings of the previous empirical studies. The last attempt to compare the proposed measures was done by Schemper and Stare (1996) [99]. More measures have been proposed since then which suggests the need for further empirical work. Next, we proposed further simulation studies to compare the alternative measures against the criteria we proposed in chapter 3. We described the simulation study design and provided the justification for choosing different parameters in the simulation studies.

In the next three chapters, we present the results of our simulation studies on the three main classes of measures and compare them systematically. We present the results of our studies on explained variation measures, explained randomness measures, and predictive accuracy measures, respectively. The measure proposed by Schemper and Kaider (1997), R^2_{SchK}, is the only measure in the "other" category that has been classified as potentially recommendable. We include the results of our investigation on this measure in chapter 7, together with the results of predictive accuracy measures.

Chapter 5

Investigation of the measures of explained variation

5.1 Introduction

This chapter studies various aspects of potentially recommendable measures in the explained variation category. The measures are R^2_{PM}, R^2_{OQF}, R^2_{XuOQ}, R^2_D, and $R^2_{Royston}$, proposed by Helland (1987) [41] and Kent & O'Quigley (1988) [49], O'Quigley and Flandre (1994) [75], Xu and O'Quigley (2001) [78], Royston & Sauerbrei (2004) [93], and Royston (2006) [88], respectively.

This chapter consists of eight sections. First, the results of simulation studies under different covariate distributions and covariate effects are presented using non-censored data. This helps us to study the expected value of the measures and investigate the impact of different covariate distributions on the measures. Second, the behaviour of the measures in different censoring mechanisms is studied in section 5.3. In section 5.4, we study the consistency and the sampling distribution of the measures, and discuss the effect of sample size on the measures. In section 5.5, the monotonicity properties of the measures are investigated. The upper bound of the measures for a range of covariate effects is illustrated in section 5.6. The behaviour of the measures in the presence of outlying observations is discussed in section 5.7. In section 5.8, the issue of model mis-specification in the context of the Cox PH model and some simulation results which elucidate the impact of model mis-specification on the measures are presented. A discussion of this chapter is presented in the last section.

5.2 Impact of covariate distribution on the measures

In this section, we evaluate the measures of explained variation under different covariate distributions and covariate effects. The aim of this section is to gain an understanding of the expected value and the spread of the sampling distribution of the measures across all covariate effects and covariate distributions in the absence of censoring. We examine the impact of censoring on the measures in section 5.3.

We present the result of simulation studies to evaluate the measures under different covariate distributions and covariate effects. In the simulation study, the survival times were simulated with four different covariate distributions as described in section 4.3.8, i.e. normal, lognormal, highly positively skewed, and highly negatively skewed distributions. The simulations were run for four covariate distributions, four covariate effects $\beta = \{0.223, 0.405,\ 0.693, 1.386\}$ representing hazard ratios of $\{1.25, 1.5, 2, 4\}$, and three sample size conditions $n = \{200, 500, 1000\}$, with 5,000 replicates in each experimental condition.

Tables 5.1 and 5.2 contain the mean and standard deviation of the sampling distribution of the measures by the covariate distribution and covariate effect, averaged over three sample size conditions. The first thing to note from table 5.1 is that the measures appear to give a good reflection of strength of association as measured by β and tend to 1 for high, but plausible, values of β. The measures generally agree with each other in normally and, to some extent, lognormally distributed covariates with the values of R^2_{OQF} and R^2_{XuOQ} slightly higher and $R^2_{Royston}$ lower. The measures differ substantially when the covariate distribution is heavily skewed across all covariate effects.

As is evident from table 5.2, the standard deviation ($S.D.$) of the measures varies across different covariate effects and distributions. Large $S.D.$ implies that the sampling distribution of the measure is more dispersed, which results in wider confidence intervals for the measure. Although the standard deviation is an informative measure of dispersion, it is difficult in this case to compare the spread of distribution across different covariate effects because the scales of the measures vary across different covariate effects. For example, for a normally distributed covariate with $\beta = 0.223$, the mean and standard deviation of the sampling distribution of R^2_{PM} are 0.031 and 0.014, respectively. When $\beta = 1.386$, however, they are equal to 0.538 and 0.034, respectively. As seen in this example, the spread of the sampling distribution, compared with the mean value, is much larger when the covariate effect, β, is equal to 0.223. Therefore, standard deviation

can be used to compare the dispersion of the measures when the covariate effects, βs, are similar (within the rows in table 5.2).

Table 5.1: Mean of the sampling distribution of explained variation measures by the covariate distribution and covariate effect across all sample size conditions, censoring=0%

Covariate Distribution	$\exp(\beta)$	R^2_{PM}	R^2_D	R^2_{OQF}	R^2_{XuOQ}	$R^2_{Royston}$
normal	1.25	0.031	0.031	0.048	0.048	0.030
	1.5	0.092	0.092	0.132	0.132	0.086
	2	0.227	0.227	0.283	0.283	0.204
	4	0.538	0.538	0.574	0.574	0.480
lognormal	1.25	0.031	0.029	0.043	0.043	0.028
	1.5	0.093	0.087	0.112	0.112	0.080
	2	0.227	0.214	0.248	0.248	0.188
	4	0.537	0.514	0.552	0.551	0.451
positively skewed	1.25	0.032	0.021	0.039	0.039	0.026
	1.5	0.093	0.059	0.105	0.105	0.069
	2	0.226	0.139	0.256	0.256	0.156
	4	0.534	0.341	0.597	0.597	0.365
negatively skewed	1.25	0.031	0.023	0.095	0.095	0.033
	1.5	0.092	0.062	0.255	0.255	0.092
	2	0.225	0.135	0.465	0.465	0.201
	4	0.533	0.291	0.728	0.728	0.428

To compare the dispersion of the measures across covariate effects (within the columns in table 5.2), it is more logical to use a measure of relative dispersion, or relative variability, than a measure of absolute dispersion or absolute variability. A better comparison of the spread of distributions can be made by using coefficient of variation ($C.V.$). Pearson [81] [1] suggested a formula for the computation of the coefficient of variation

$$C.V. = \frac{S.D.}{Mean}.$$

This is one way of standardising the dispersion of the measures to improve comparability across covariate effects. The coefficient of variation, $C.V.$ is only a good measure of dispersion when $Mean > 0$. Table 5.3 shows the average coefficient of variation of the measures across three sample size conditions in the non-censored condition, expressed as percentages. The spread of the distribution of measures is similar in the normally distributed covariate. The distribution of $R^2_{PM}, R^2_D, R^2_{OQF}$, and R^2_{XuOQ} become relatively more dispersed as the skewness of the covariate becomes larger. The relative spread of distribution remains unchanged for $R^2_{Royston}$ in different covariate distributions. Finally, as

[1] Reference taken from Paul L. Boynton, "The Coefficient of Variation as a Tool in Educational Practice", Peabody Journal of Education (1934), 11(5), 216-224.

Table 5.2: Standard deviation of the sampling distribution of explained variation measures by the covariate distribution and covariate effect across all sample size conditions, censoring=0%

Covariate Distribution	exp(β)	R^2_{PM}	R^2_D	R^2_{OQF}	R^2_{XuOQ}	$R^2_{Royston}$
normal	1.25	0.014	0.014	0.022	0.022	0.013
	1.5	0.024	0.024	0.033	0.033	0.022
	2	0.034	0.034	0.040	0.040	0.030
	4	0.034	0.034	0.037	0.037	0.033
lognormal	1.25	0.015	0.014	0.019	0.019	0.013
	1.5	0.024	0.024	0.029	0.029	0.020
	2	0.035	0.034	0.040	0.040	0.028
	4	0.038	0.037	0.045	0.057	0.034
positively skewed	1.25	0.016	0.012	0.020	0.020	0.011
	1.5	0.028	0.020	0.040	0.040	0.018
	2	0.045	0.031	0.074	0.074	0.027
	4	0.055	0.043	0.074	0.074	0.037
negatively skewed	1.25	0.016	0.012	0.078	0.078	0.016
	1.5	0.031	0.020	0.100	0.100	0.026
	2	0.051	0.029	0.091	0.091	0.035
	4	0.061	0.039	0.055	0.055	0.039

the covariate effect becomes smaller, the spread of the distribution of measures becomes larger. Some of the findings are highlighted below for each measure in this category.

5.2.1 Helland (1987) and Kent & O'Quigley (1988) measure - R^2_{PM}

The mean of the sampling distribution of R^2_{PM} varies from 0.031 to 0.538 for different covariate effects. The measure is independent of the shape of the covariate distribution. The dispersion of the measure decreases as the covariate effect increases. For example, the coefficient of variation decreases from 43.4% for $\beta = 0.223$ (hazard ratio of 1.25) to 6% for $\beta = 1.386$ (hazard ratio of 4) when the covariate distribution is normal. There is a similar pattern in other covariate distributions, but with bigger dispersion in skewed covariate distributions.

5.2.2 Royston and Sauerbrei measure (2004) - R^2_D

The mean and dispersion of the sampling distribution of R^2_D are similar to those of R^2_{PM} for normally distributed covariates. The expected value of the measure decreases as the covariate distribution becomes asymmetrical. The larger the skewness of the covariate distribution, the larger the decrease. For example, for $\beta = 1.386$ (hazards ratio of 4), the expected value of the measure decreases from 0.536 in the normally distributed covariate

Table 5.3: Coefficient of variation of explained variation measures by the covariate distribution and covariate effect, expressed as %. Table entries are the average across all combinations of sample sizes, censoring=0%.

Covariate Distribution	exp(β)	R^2_{PM}	R^2_D	R^2_{OQF}	R^2_{XuOQ}	$R^2_{Royston}$
normal	1.25	43.4	43.5	42.6	42.6	42.6
	1.5	24.4	24.5	23.7	23.7	24.0
	2	14.1	14.1	13.5	13.5	14.0
	4	6.0	6.1	6.0	6.0	6.5
lognormal	1.25	43.7	44.9	41.0	41.0	42.0
	1.5	24.9	25.7	23.7	23.7	23.8
	2	14.7	15.2	14.9	14.9	14.3
	4	6.6	6.7	7.7	10.3	7.1
positively skewed	1.25	46.1	53.4	46.5	46.5	42.2
	1.5	28.4	32.8	35.3	35.3	25.2
	2	18.8	21.4	27.3	27.3	16.5
	4	9.8	12.0	11.9	11.9	9.5
negatively skewed	1.25	49.0	51.3	79.1	79.1	47.3
	1.5	31.7	30.8	37.9	37.9	27.5
	2	21.5	20.2	18.7	18.7	16.7
	4	10.9	12.8	7.2	7.2	8.6

to 0.514 when the skewness is equal to 1, that is lognormal covariate, and to 0.341 when the skewness is equal to 2.8.

This reflects the impact of non-normality of the covariate or the prognostic index (PI) on the D measure [93], which was reported by Royston and Sauerbrei (2004) [93]. They showed that on average non-normality of the PI appears to reduce the D measure . To compute D, first the Cox PH model is fitted. Then the prognostic index of the model, $\beta' X$, is transform to give standard normal order rank statistics (rankits - formed using Blom's approximation [93]). The rankits are multiplied by a factor of $\sqrt{8/\pi}$ to give Z_i ($i = 1, .., n$ subjects). Finally a Cox PH model is fitted to these values; D is the coefficient of Z, say σ^*, from this second model. Royston and Sauerbrei (2004) [93] showed that D most accurately measures separation of survival curves when the underlying prognostic index values, $\beta' x_i$, are normally distributed. The regression on the Z in the second model is then linear and σ^* is an approximately unbiased estimate of σ. They explained that when the $\beta' x_i$ are not normally distributed, linearity in the second model breaks down [93]. D still measures separation because σ^* in the second model still estimates σ, but with bias.

5.2.3 O'Quigley and Flandre measure (1994) - R^2_{OQF}

This measure and R^2_{XuOQ} are identical in non-censored data. They seem to have higher values than the other measures in this category. The mean of the sampling distribution is about 0.048 for $\beta = 0.223$, and increases to 0.574 for $\beta = 1.386$. This measure is also influenced by the covariate distribution. The impact of covariate distribution on this measure seems to depend on the strength of the relationship and the skewness of covariate distribution. The dispersion of the measure increases as the distribution of the covariate in the model becomes skewed.

5.2.4 Xu and O'Quigley measure (2001) - R^2_{XuOQ}

This measure is identical to O'Quigley and Flandre's measure (1994), R^2_{OQF}, in non-censored data. Thus, it possesses the same properties as the O'Quigley and Flandre measure (1994).

5.2.5 Royston measure (2006) - $R^2_{Royston}$

The mean of the sampling distribution of this measure varies from 0.026 to 0.480, depending on the strength of the relationship and the skewness of the covariate. The dispersion of this measure decreases with increasing covariate effect, and the covariate skewness has less impact on this measure compared with its impact on the others.

5.3 Impact of censoring on the measures

The impact of censoring was studied by considering two types of censoring mechanisms, type I, known as administrative censoring, and random censoring, and four censoring proportions. The mechanisms applied for generating each censoring type was explained in section 4.3.4. Simulations were run for two types of censoring mechanisms, four censoring proportions, 0%, 20%, 50%, and 80%, four covariate effects $\beta = \{0.223, 0.405,\ 0.693, 1.386\}$ representing hazard ratios of $\{1.25, 1.5, 2, 4\}$, and three sample size conditions $n = \{200, 500, 1000\}$, with 5,000 replicates in each experimental condition.

Table 5.4 shows the average percentage difference of measures to the expected value of corresponding non-censored data by the covariate distribution and censoring proportion. The entries in the table are the average across two censoring types, four covariate effects,

and three sample size conditions, as outlined in section 4.3. Table 5.5 shows the relative dispersion of the measures expressed as *C.V.*, averaged over the same experimental conditions.

Furthermore, table 5.6 displays the impact of censoring type on the expected value and dispersion of the measures by covariate distributions. The figures in this table are the average across four censoring proportions, four covariate effects, and three sample size conditions. Detailed simulation results are presented in Appendix A. The tables in Appendix A show the impact of censoring by the covariate distribution, censoring type, and censoring proportion in a similar way to table 5.6.

Since one of the aims of this study was to make practical comparisons of the measures, we translate the figures in table 5.4 into 4 categories each representing the extent of censoring. The categories are:

1) almost no effect: the average percentage change in the expected value of the measure is $0\% - 9\%$ compared to that of non-censored data.

2) slight effect: the average percentage change in the expected value of the measure is $10\% - 19\%$ compared to that of non-censored data.

3) moderate effect: the average percentage change in the expected value of the measure is $20\% - 49\%$ compared to that of non-censored data.

4) large effect: the average percentage change in the expected value of the measure is more than 50% compared to that of non-censored data.

This classification helps us to interpret the results and easily compare the measures. The impact of censoring on each measure is summarised in the following sections.

5.3.1 Helland (1987) and Kent & O'Quigley (1988) measure - R^2_{PM}

The amount of censoring has the least impact on this measure among the explained variation measures. As it appears from table 5.4, the measure increases slightly with the amount of censoring. For instance, with 80% censoring and normally distributed covariates, the measure is on average 6.3% higher than the value of the measure with the corresponding non-censored data. The spread of the sampling distribution of this measure also increases as the amount of censoring increases for all covariate distributions. Table 5.6

Table 5.4: The average percentage difference from the expected value of the measures in the corresponding non-censored data by the covariate distribution and censoring proportion.

Covariate Distribution	% Censored	R^2_{PM}	R^2_D	R^2_{OQF}	R^2_{XuOQ}	$R^2_{Royston}$
normal	20	0.4	0.4	2.3	1.8	5.3
	50	1.6	1.6	6.8	5.6	13.1
	80	6.3	6.4	16.0	14.4	26.9
lognormal	20	0.2	4.9	3.5	2.3	10.2
	50	1.0	13.5	11.0	8.1	28.0
	80	4.0	28.1	25.1	20.6	58.1
positively skewed	20	0.1	13.1	2.9	1.5	16.0
	50	0.5	40.2	11.3	7.2	50.1
	80	2.2	88.9	30.6	23.6	115.9
negatively skewed	20	1.3	-10.6	-12.3	-9.3	-10.6
	50	4.5	-19.7	-19.5	-14.7	-21.1
	80	16.7	-21.3	-18.8	-14.3	-24.5

Table 5.5: Coefficient of variation of explained variation measures by the covariate distribution and censoring proportion, expressed as %.

Covariate Distribution	% Censored	R^2_{PM}	R^2_D	R^2_{OQF}	R^2_{XuOQ}	$R^2_{Royston}$
normal	20	23.6	23.7	23.1	23.4	23.9
	50	28.4	28.5	27.7	31.8	29.2
	80	40.1	40.3	39.1	57.3	42.1
lognormal	20	23.5	24.6	22.9	23.1	23.6
	50	27.1	28.8	26.4	27.9	28.3
	80	36.4	39.1	35.8	51.8	39.7
positively skewed	20	26.3	31.0	47.2	48.2	24.7
	50	28.5	34.6	35.5	43.3	28.5
	80	34.6	42.9	37.9	58.4	37.5
negatively skewed	20	31.8	31.6	34.0	37.2	27.8
	50	39.3	39.1	39.4	50.3	34.6
	80	57.2	56.4	53.5	71.4	50.6

shows that the mean and relative dispersion of the sampling distribution of this measure are similar in both censoring types.

5.3.2 Royston and Sauerbrei measure (2004) - R^2_D

The impact of censoring on this measure depends on the covariate distribution. In a model whose covariate or prognostic index distribution is positively skewed, the measure increases as the amount of censoring increases. In contrast, when the covariate is negatively skewed, the measure decreases as the amount of censoring increases. The impact of censoring becomes larger as the covariate becomes more skewed. The spread of sampling distribution of this measure in censored data is similar to that of R^2_{PM} with the exception

Table 5.6: Summary performance of explained variation measures by the covariate distribution and censoring mechanism.

Measure	Covariate Distribution	Random Censoring		Type I Censoring	
		Average % Difference	*C.V.*	Average % Difference	*C.V.*
R^2_{PM}	normal	2.9	30.9	2.6	30.4
	lognormal	2.0	29.4	1.5	28.7
	positively skewed	1.3	30.1	0.6	29.4
	negatively skewed	7.1	42.4	7.9	43.2
R^2_D	normal	2.9	31.1	2.7	30.6
	lognormal	14.4	31.2	16.7	30.5
	positively skewed	43.6	36.9	51.2	35.4
	negatively skewed	-14.6	42.5	-19.8	42.2
R^2_{OQF}	normal	8.1	30.3	8.6	29.6
	lognormal	13.1	28.7	13.3	28.0
	positively skewed	16.1	37.4	13.8	43.0
	negatively skewed	-13.8	44.0	-19.8	40.7
R^2_{XuOQ}	normal	5.9	45.3	8.6	29.6
	lognormal	7.4	40.6	13.3	28.0
	positively skewed	7.8	56.9	13.8	43.0
	negatively skewed	-5.7	65.3	-19.8	40.7
$R^2_{Royston}$	normal	13.5	31.9	16.7	31.5
	lognormal	28.8	30.8	35.4	30.3
	positively skewed	55.1	30.7	66.3	29.8
	negatively skewed	-16.5	37.7	-20.9	37.6

of positively skewed distribution, which is higher than that of R^2_{PM}. Table 5.6 shows that type I censoring has more impact on the expected value of the measure than random censoring. However, the spread of the sampling distribution seems to be similar under both random and type I censoring.

5.3.3 O'Quigley and Flandre measure (1994) - R^2_{OQF}

The impact of censoring on this measure also depends on the skewness of the covariate distribution. Table 5.4 makes it clear that while the measure decreases as the amount of censoring increases in negatively skewed covariates, it increases as the amount of censoring decreases in positively skewed covariates. Table 5.5 shows that the spread of the sampling distribution increases as both censoring and the covariate skewness increase. For example, in positively skewed covariates, the measure is on average 16.1% higher under random censoring conditions compared with the value of the measure in the corresponding non-censored data. However, it is on average 13.8% lower compared with the expected value of the measure in the corresponding non-censored data if the covariate is negatively

distributed.

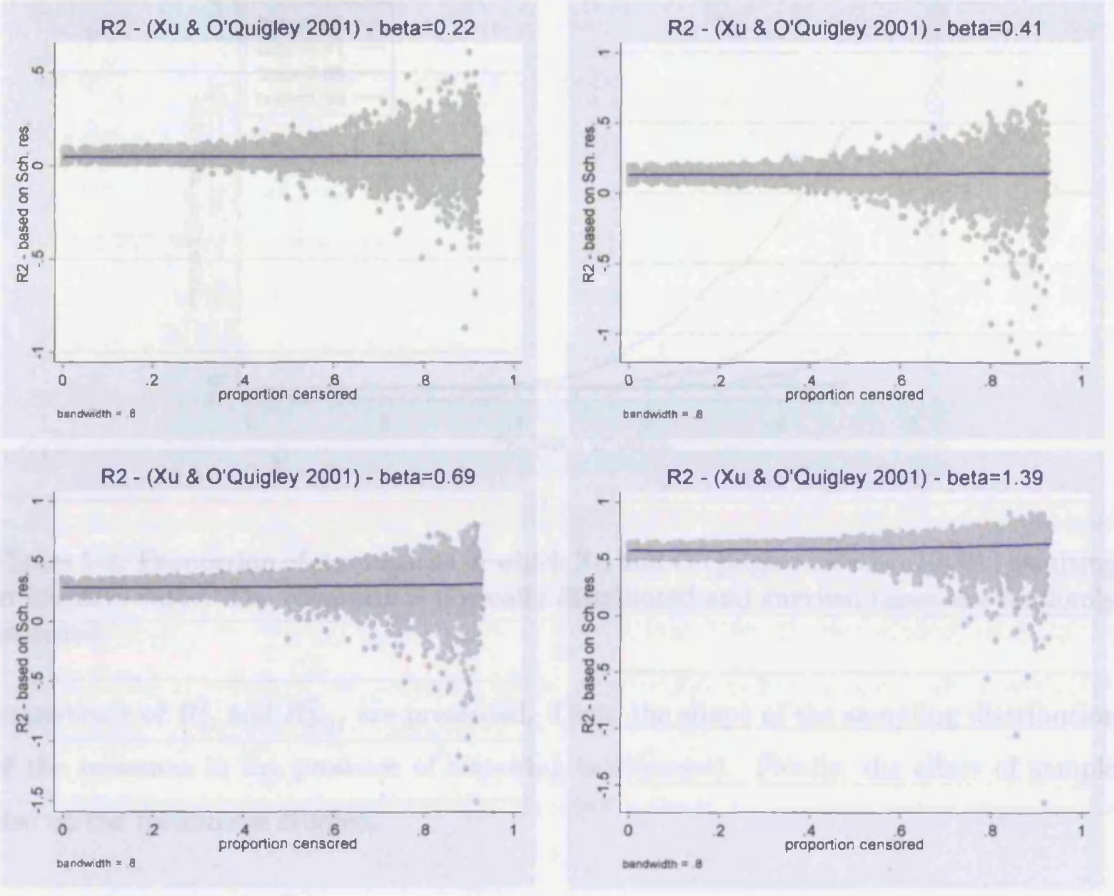
5.3.4 Xu and O'Quigley measure (2001) - R^2_{XuOQ}

This measure is a modification of O'Quigley and Flandre's measure (1994) [75], R^2_{OQF}. As discussed in Xu (1996) [115] and Xu and O'Quigley (2001) [78], to eliminate the asymptotic dependence of O'Quigley and Flandre's measure (1994) [75] on censoring, it is necessary to weight the squared Schoenfeld residual in O'Quigley and Flandre's measure (1994) [75] by increments of any consistent estimate of the marginal failure time distribution function. Xu and O'Quigley (2006) [79] explained that the practical impact of this weighting on numerical values would typically be small. The results in table 5.4 emphasise this theory. However, the results of the simulation study show that censoring still has a minor effect on this measure and the weighting scheme has not eliminated its impact completely.

The weighting scheme diminishes the impact of censoring on this measure compared with its impact on O'Quigley and Flandre's measure (1994) [75]. Nevertheless, the spread of the sampling distribution of this measure increases dramatically as the censoring proportion becomes larger. For example, the *C.V.* of Xu and O'Quigley's measure (2001) [78] in the normally distributed covariate is on average 57.3, where that of O'Quigley and Flandre's measure (1994) [75] is on average 39.1. Table 5.6 shows that this measure is identical to O'Quigley and Flandre's measure (1994) [75] in type I censoring. Random censoring has less impact on this measure compared with O'Quigley and Flandre's measure (1994) [75] in all covariate distributions.

Further assessment of the simulation results revealed an undesirable impact of censoring on this measure. Figure 5-1 demonstrates this finding in more detail. The figure consists of four graphs one for each covariate effect. In the graphs, the dots represent the estimates of this measure in each replicate and the solid line is the expected value of the measure when the covariate is normally distributed from 0% to 90% censoring. As it is evident, the expected value of the measure is consistent as the amount of censoring increases across four covariate effects. But the measure cannot be guaranteed to be non-negative. In fact, as figure 5-2 demonstrates, the chance that the measure leads to a negative value increases as the amount of censoring goes up.

Figure 5-1: The expected value (solid line) of Xu and O'Quigley measure (2001) by the censoring proportion when the covariate is normally distributed, random censoring condition, and sample size=1000, Dots are the estimates of the measure in each replicate.



5.3.5 Royston measure (2006) - $R^2_{Royston}$

Among the measures in this category, $R^2_{Royston}$'s performance is the worst with regard to the impact of censoring. Table 5.4 reveals that the censoring has the biggest impact on this measure, compared with other measures in this category, in all censoring proportions and covariate distributions. Table 5.6 also shows that type I censoring has more impact on this measure than random censoring.

5.4 Consistency, distributional shape, and sample size effect

In this section, we discuss the consistency and the shape of the sampling distribution of the measures as well as the effect of sample size on them. First, the characteristic of a consistent estimator together with the results of the simulation study to investigate the

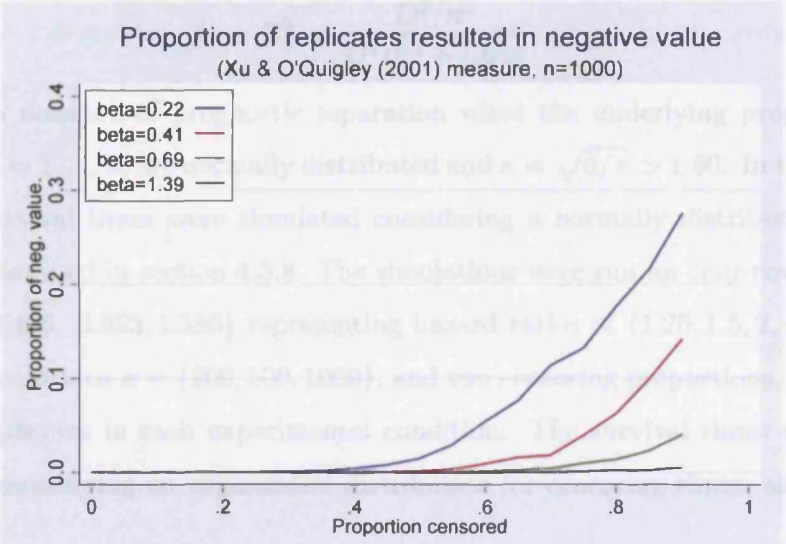


Figure 5-2: Proportion of simulations in which Xu and O'Quigley measure (2001) resulting in negative value. The covariate is normally distributed and survival times are randomly censored.

consistency of R^2_D and R^2_{PM} are presented. Then, the shape of the sampling distribution of the measures in the presence of censoring is discussed. Finally, the effect of sample size on the measure is studied.

5.4.1 Consistency of the measures

An important characteristic of estimators is consistency. Formally, $\widehat{\theta}$ is a consistent estimator of the parameter θ if and only if $Lim_{n\to\infty} \Pr\left(\left|\widehat{\theta} - \theta\right| < \epsilon\right) = 1$ for every $\epsilon > 0$ (Mood et al (1974) [70], page 295). Less formally, a consistent estimator is one for which the probability that it is arbitrarily close to the parameter converges to 1 as the sample size, n, increases without bound. A consistent estimator is not necessarily unbiased in finite samples, but as the sample becomes larger and larger, the estimator gets closer and closer in value to the parameter of interest (Mood et al (1974) [70], page 295). If the measures studied here are consistent under different censoring proportions, their bias should decrease toward 0 and the spread of their sampling distributions should become smaller and smaller as the sample size increases. Therefore, bias and MSE of the estimators should be investigated.

In this section, we explore the consistency of the measure proposed by Royston and Sauerbrei (2004) [93] through some simulation studies. For the Cox PH model, the

measure is defined as

$$R_D^2 = \frac{D^2/\kappa^2}{D^2/\kappa^2 + 1.645}$$

where D is a measure of prognostic separation when the underlying prognostic index values $\beta' X_i$, $i = 1, ..., n$, are normally distributed and $\kappa = \sqrt{8/\pi} \simeq 1.60$. In the simulation study, the survival times were simulated considering a normally distributed covariate, $N(0, 1)$, as described in section 4.3.8. The simulations were run for four covariate effects $\beta = \{0.223, 0.405,\ 0.693, 1.386\}$ representing hazard ratios of $\{1.25, 1.5, 2, 4\}$, and three sample size conditions $n = \{200, 500, 1000\}$, and two censoring proportions, 0% and 80%, with 5000 replicates in each experimental condition. The survival times are randomly censored by considering an exponential distribution for censoring times, as described in section 4.3.4.

In the above setting, a model with normally distributed covariate, D^2/κ^2 is, by definition, the variance of the prognostic index, $\beta' X_i$, of the model (Royston and Sauerbrei (2004) [93]). If $X \sim N(0, 1)$ then $\beta' X_i \sim N(0, \beta^2)$; therefore, the population value of the R_D^2 for different values of the covariate effect, $\beta = \{0.223, 0.405,\ 0.693, 1.386\}$, are 0.029, 0.091, 0.226, and 0.539, respectively.

Table 5.7 displays the results of the simulation study in terms of the estimated bias and the estimated $\sqrt{MSE}$ of the $\widehat{R}_D^2$. The bias is defined as

$$Bias = \overline{R}_D^2 - R_D^2$$

where $\overline{R}_D^2$ is the empirical mean across $5,000$ sampling repetitions used in a given experimental condition

$$\overline{R}_D^2 = \frac{\sum_{i=1}^{5000} R_{D_i}^2}{5000}.$$

The empirical root mean squared error, $\sqrt{MSE}$, is defined as

$$MSE = Bias^2 - SE^2$$

where

$$SE = \left(\frac{\sum_{i=1}^{5000} \left(R_{D_i}^2 - \overline{R}_D^2 \right)^2}{5000} \right)^{1/2}.$$

Table 5.7 shows that the sample estimate $\widehat{R}^2_D$ is unbiased in non-censored data for all cases considered. In the censored data, the sample estimate has a slight positive bias for the small covariate case, i.e. $\beta = \{0.223, 0.405\}$, when the sample size is small, i.e. $n = 200$; otherwise, the sample estimate is unbiased. In terms of MSE, the sample estimate has larger MSE in censored data compared to the non-censored one, as expected. As it is evident from the table, the MSE and Bias of the sample estimate $\widehat{R}^2_D$ becomes smaller as the sample size, n, increases in both censored and non-censored data. We, therefore, can conclude that the sample estimate $\widehat{R}^2_D$ is a consistent estimator of R^2_D in normally distributed covariates.

Table 5.7: Summary of the estimated bias and root mean squared error (RMSE) of the estimator of Royston and Sauerbrei measure (2004). Normally distributed covariate and randomly censored data.

R^2_D	n	0% censoring		80% censoring	
		Bias	$\sqrt{MSE}$	Bias	$\sqrt{MSE}$
0.029	200	0.003	0.020	0.014	0.047
	500	0.001	0.012	0.006	0.027
	1000	0.001	0.008	0.003	0.019
0.091	200	0.003	0.033	0.012	0.069
	500	0.001	0.021	0.005	0.043
	1000	0.001	0.014	0.002	0.030
0.226	200	0.001	0.046	0.006	0.088
	500	0.001	0.029	0.004	0.056
	1000	0.000	0.021	0.002	0.039
0.539	200	-0.005	0.047	-0.002	0.080
	500	-0.002	0.030	0.000	0.050
	1000	-0.001	0.021	0.000	0.035

Similar analysis was performed for the R^2_{PM} to investigate the consistency of this measure. The results were very similar to those of R^2_D, thus the same conclusion can be drawn on R^2_{PM}. In section 5.2, we showed that the mean and dispersion of the sampling distribution of R^2_D are also similar to those of R^2_{PM} for a normally distributed covariate.

The consistency of other measures in this category were studied before. The measures proposed by O'Quigley and Flandre (1994) [75], R^2_{OQF}, is a consistent estimator of the population value, $\mathbf{R}^2_{OQF}$ as expressed in equation 2.29, in the absence of censoring ([75] and [115]). However, Xu (1996) [115] showed that R^2_{OQF} depends upon censoring even asymptotically in the presence of censoring. Xu (1996) [115] introduced R^2_{XuOQ} and analytically established its consistency as an estimator of the population value, $\mathbf{R}^2_{OQF}$. The measure proposed by Royston (2006) [88] is a transformation of the explained randomness measure proposed by O'Quigley et al (2005) [80], ρ^2_k. As pointed out by O'Quigley et al

(2005) [80], ρ_k^2 is consistent in the absence of censoring. But, it converges to a different population quantity in censored data. More discussion on the consistency of ρ_k^2, along with other explained randomness measures, is presented in section 6.4 of next chapter. In the next section, we present the sampling distribution of the measures for different sample sizes, covariate effects, and censoring proportions.

5.4.2 Sampling distribution of the measures

Generally, simulation results show that for small sample sizes and small covariate effects, the sampling distributions of the estimators of explained variation measures exhibit considerable skewness, particularly when censoring is more than 50%. Figure 5-3 is presented as an example to depict the distributional properties of the measures. The figure shows the sampling distribution of Royston and Sauerbrei's measure (2004) [93], R_D^2, by the censoring proportion, covariate effect, and sample size. The smooth curves in the figure are the kernel density estimates in each experimental condition. The survival times are randomly censored by considering an exponential distribution for censoring times, as described in section 4.3.4. The covariate, or prognostic index in the case of multiple regression, of the model is normally distributed and the number of replicates are $5,000$ in each experimental condition.

As seen in figure 5-3, more symmetry is evident as the covariate effect, β, and sample size, n, become larger. By the time n attains $1,000$, however, virtually all distributions are approximately bell shaped in small to moderate censoring, i.e. when the censoring proportion is not more than 50%. The positive skewness in all distributions is quite evident when censoring is heavy and sample size is small. We explored the sampling distribution of other measures in this category with the same experimental conditions. The shape of the sampling distribution of the measures follows a similar pattern, except R_{XuOQ}^2 which results in negative values as censoring increases, as explained in section 5.3.4.

We can also crudely explore the consistency of the estimators graphically over the range of n in this study. Sampling distributions of consistent estimators should tend towards a spike over the parameter of interest as n becomes ever larger. Intuitively, this means that the sampling distribution of a consistent estimator becomes more and more concentrated on the parameter of interest as n becomes ever larger. All distributions in figure 5-3 appear to exhibit this tendency, although some more so than the others,

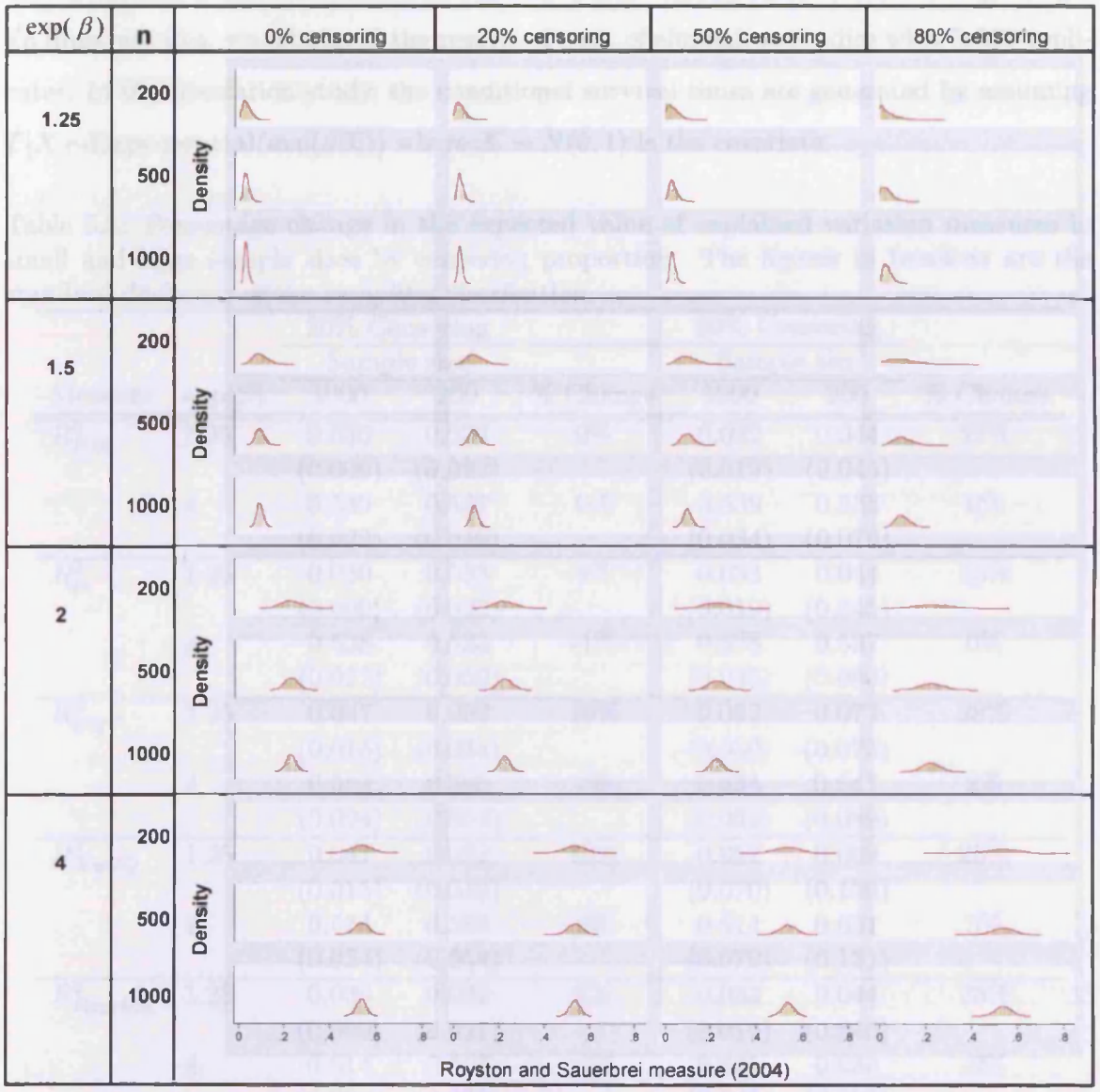


Figure 5-3: Sampling distributions of Royston and Sauerbrei measure (2004) by the covariate effect, sample size, and censoring proportions in the normally distributed covariate and random censoring condition.

depending on the censoring proportion.

5.4.3 Impact of sample size on the measures

The results of all simulation studies show that in all explained variation measures sample size has an effect when the covariate effect is small and the amount of censoring is high. To illustrate this, we tabulated the results of a set of simulation studies with 5,000 replicates. In the simulation study, the conditional survival times are generated by assuming $T|X \sim$**Exponential**$(\exp(\beta X))$ where $X \sim N(0,1)$ is the covariate.

Table 5.8: Percentage change in the expected value of explained variation measures in small and large sample sizes by censoring proportion. The figures in brackets are the standard deviation of the sampling distribution.

		20% Censoring			80% Censoring		
		Sample size			Sample size		
Measure	$\exp(\beta)$	1000	200	% Change	1000	200	% Change
R^2_{PM}	1.25	0.030	0.033	9%	0.032	0.044	27%
		(0.009)	(0.022)		(0.019)	(0.045)	
	4	0.539	0.537	0%	0.539	0.538	0%
		(0.022)	(0.049)		(0.034)	(0.076)	
R^2_D	1.25	0.030	0.033	9%	0.033	0.044	25%
		(0.009)	(0.022)		(0.019)	(0.045)	
	4	0.538	0.534	-1%	0.538	0.537	0%
		(0.022)	(0.050)		(0.035)	(0.080)	
R^2_{OQF}	1.25	0.047	0.052	10%	0.052	0.072	28%
		(0.015)	(0.034)		(0.030)	(0.073)	
	4	0.584	0.590	1%	0.634	0.647	2%
		(0.024)	(0.054)		(0.039)	(0.089)	
R^2_{XuOQ}	1.25	0.047	0.052	10%	0.052	0.069	25%
		(0.015)	(0.035)		(0.070)	(0.120)	
	4	0.574	0.583	2%	0.614	0.631	3%
		(0.024)	(0.054)		(0.079)	(0.131)	
$R^2_{Royston}$	1.25	0.030	0.032	6%	0.033	0.044	25%
		(0.009)	(0.021)		(0.019)	(0.046)	
	4	0.514	0.507	-1%	0.641	0.630	-2%
		(0.024)	(0.053)		(0.045)	(0.100)	

Random non-informative right censoring was generated as described in section 4.3.4. Table 5.8 shows that the measures increase when both sample size and the covariate effect are small, i.e. $n = 200$ and $\exp(\beta) = 1.25$, and the amount of censoring is high, i.e. 80%. This pattern was observed in other simulation studies when we considered skewed covariates and a different censoring mechanism, i.e. type I censoring.

5.5 Monotonicity properties of the proposed measures

In this section, the parameter and number of variables monotonicity properties of the explained variation measures are investigated. In chapter 3, these two properties were considered essential for a suitable measure of predictive ability. Parameter monotonicity means that the expected value of predictive ability measures should not decrease as the absolute value of covariate effect, β, in the model increases. The number of variables monotonicity means that the expected value of a suitable measure of predictive ability should not decrease by adding new covariates. This section is divided into two parts, describing the two monotonicity properties separately.

5.5.1 Parameter monotonicity

The parameter monotonicity of R^2_{OQF} and R^2_{XuOQ} has been established analytically by O'Quigley and Flandre (1994) [75] and Xu (1996) [115]. Furthermore, R^2_{PM} satisfies this property since it is a monotonic function of $|\beta|$. Equation 2.24 can be written as

$$\begin{aligned} R^2_{PM} &= \frac{\mathbf{Var_X}(\boldsymbol{\beta}'\mathbf{x})}{\mathbf{Var_X}(\boldsymbol{\beta}'\mathbf{x}) + \boldsymbol{\pi^2/6}} \\ &= 1 - \frac{\boldsymbol{\pi^2/6}}{\boldsymbol{\beta^2}\mathbf{Var_X}(\mathbf{x}) + \boldsymbol{\pi^2/6}} \end{aligned}$$

which is an increasing function of the covariate effect, β. Similarly, the measure proposed by Royston and Sauerbrei (2004) [93], R^2_D, can be written as

$$\begin{aligned} R^2_D &= \frac{D^2/\kappa^2}{D^2/\kappa^2 + \sigma^2_\varepsilon} \\ &= 1 - \frac{\sigma^2_\varepsilon}{D^2/\kappa^2 + \sigma^2_\varepsilon} \end{aligned}$$

where $\sigma^2_\varepsilon = \pi^2/6$ for the Cox PH model. This shows that R^2_D is an increasing function of the D measure ([93]) which is a monotonic function of $|\beta|$ (Royston and Sauerbrei (2004) [93]) when the prognostic index of the model, $\boldsymbol{\beta}'\mathbf{x}$, is assumed to be normally distributed. The measure proposed by Royston (2006) [88], $R^2_{Royston}$, inherits ρ^2_k properties of which parameter monotonicity is one (section 6.6 or O'Quigley et al (2005) [80]).

Furthermore, the simulation results displayed in table 5.1 show that the expected value of the measures increase as the covariate effect becomes stronger in all covariate distributions. Moreover, the results of another simulation study, performed to investigate

the upper bound of the measures for a wide range of covariate effects, presented in section 5.6, show that the measures are an increasing function of the covariate effect. Thus, all of them satisfy this property.

5.5.2 Number of variables monotonicity

As it was described in chapter 3, an appropriate measure of predictive ability should not decrease as new prognostic factors, i.e. covariates, are included in the model. The number of variables monotonicity means that the expected value of a measure of predictive ability should not decrease by adding new covariates. In this section, a further simulation study was carried out to investigate the impact of adding new but independent covariates to the model. The simulation study was carried out for four covariate effects and two censoring proportions. The sample size was 500, and 2,000 replicates were generated for each experimental conditions. In the simulation, the distribution of survival time is generated using the algorithm outlined in section 4.3.9 by assuming only one covariate that is normally distributed. Then, two new covariates were generated independently and the following models were fitted in each replicate: Model I with only the dependent covariate; Model II with the dependent covariate and one independent covariate; and Model III with the dependent covariate and two independent covariates.

Table 5.9 displays the differences in the expected values of the measures after fitting models II and III compared to model I. The table shows that the expectation of the measures do not decrease as new but independent covariates are included in the model in both non-censored and censored conditions.

Table 5.10 displays the results of the simulation study, summarised in terms of proportions. The entries in the table are the proportion of simulations in which the measure decreased after adding one or two independent covariates to the model. For example, when $\exp(\beta) = 1.25$ ($\beta = 0.22$) and the amount of censoring is 0%, in about 10% of 2000 simulations the value of R^2_{PM} decreased after adding one new independent covariate. The proportion of simulations in which the measure decreased fell to 3% after two independent covariates were added to the model.

As seen, $R^2_{Royston}$ always increases after adding new covariate to the model. This measure is based on the likelihood function which always increases by adding a new covariate to the model, regardless of whether the covariate is related to the outcome. It is also clear that the performance of R^2_D, R^2_{OQF}, and R^2_{XuOQ} is similar in non-censored

Table 5.9: Mean difference in the expected value of the measures after adding one or two independent covariates to the model in 2,000 simulations, normally distributed covariates.

Measure	exp(β)	Model II		Model III	
		Mean difference to model I		Mean difference to model I	
		0% censoring	80% censoring	0% censoring	80% censoring
R^2_{PM}	1.25	0.001	0.006	0.003	0.012
	1.5	0.001	0.005	0.003	0.011
	2	0.001	0.005	0.003	0.009
	4	0.001	0.003	0.002	0.006
R^2_D	1.25	0.001	0.006	0.003	0.012
	1.5	0.001	0.005	0.003	0.011
	2	0.001	0.005	0.003	0.009
	4	0.001	0.003	0.002	0.006
R^2_{OQF}	1.25	0.002	0.009	0.004	0.018
	1.5	0.002	0.007	0.003	0.015
	2	0.001	0.005	0.002	0.010
	4	0.001	0.002	0.001	0.004
R^2_{XuOQ}	1.25	0.002	0.009	0.004	0.019
	1.5	0.002	0.008	0.003	0.016
	2	0.001	0.004	0.002	0.010
	4	0.001	0.002	0.001	0.005
$R^2_{Royston}$	1.25	0.001	0.006	0.002	0.012
	1.5	0.001	0.006	0.002	0.012
	2	0.001	0.005	0.002	0.011
	4	0.001	0.003	0.002	0.006

data. But, R^2_{XuOQ} performs worse in censored conditions, i.e. the chance that the measure goes down after adding a new independent covariate to the model is more than the other measures.

5.6 Upper bound of the measures

In this section, more simulation studies are carried out to investigate the upper bound of explained variation measures. In the simulations, the predictor X is normally distributed and the distribution of the conditional survival times are exponentially distributed, i.e. $T|X \sim$**Exponential**(exp(βX)). Random non-informative right censoring are generated by considering an exponential distribution for censoring times as described in section 4.3.4. The simulations are carried out for a wide range of covariate effects from small to large, but reasonable, values with 2,000 replicates in each experimental condition. Figure 5-4 displays the expected value of the measures from $\beta = 0.22$ (exp(β) = 1.25) to $\beta = 5.55$ (exp(β) = 256) for 0% and 50% censoring. In both censoring conditions, the expected values of the measures increase with the covariate effect, and they reach values

Table 5.10: Proportion decrease in measures after adding one or two independent covariates to the model in 2,000 simulations, normally distributed covariates.

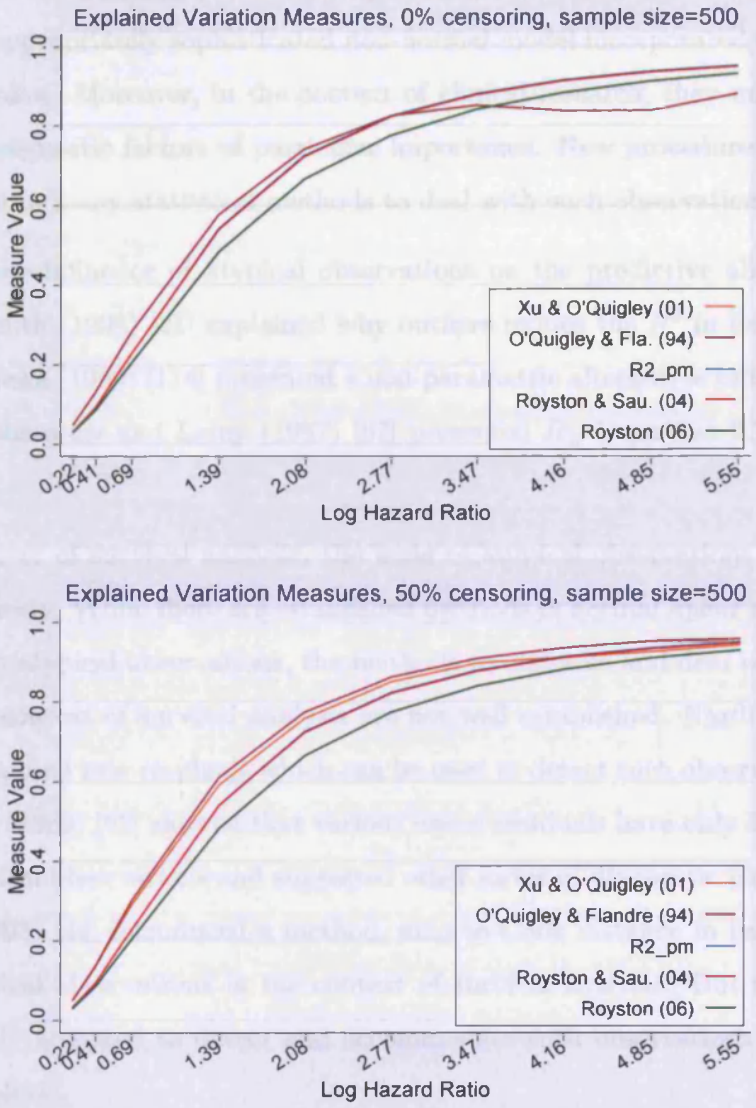
Measure	exp(β)	Model II		Model III	
		prop. decreased to model I		prop. decreased to model I	
		0% censoring	80% censoring	0% censoring	80% censoring
R^2_{PM}	1.25	0.10	0.07	0.03	0.02
	1.5	0.15	0.12	0.06	0.04
	2	0.18	0.19	0.09	0.10
	4	0.20	0.27	0.09	0.18
R^2_D	1.25	0.18	0.16	0.09	0.05
	1.5	0.24	0.20	0.13	0.09
	2	0.27	0.26	0.17	0.15
	4	0.30	0.34	0.21	0.23
R^2_{OQF}	1.25	0.17	0.16	0.07	0.06
	1.5	0.20	0.23	0.11	0.12
	2	0.25	0.29	0.16	0.20
	4	0.30	0.34	0.22	0.26
R^2_{XuOQ}	1.25	0.17	0.40	0.07	0.32
	1.5	0.20	0.40	0.11	0.33
	2	0.25	0.42	0.16	0.36
	4	0.30	0.42	0.22	0.35
$R^2_{Royston}$	1.25	0.00	0.00	0.00	0.00
	1.5	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00

near, but less than, 1. This suggests that, first, the measures have the upper limit of 1 and, second, they will lead to high values for large, but reasonable, covariate effects, i.e. they reach values more than 0.80.

5.7 Robustness of the measures

In this section, we study the impact of extreme and outlier observations on the explained variation measures. Barnett and Lewis (1994) [9] defined an outlier as an observation (or a set of observations) which appears to be inconsistent with the remainder of that set of data. Barnett and Lewis (1994) [9] made a clear distinction between outliers and extreme observations and argued that they are not coincident concepts. An outlier observation may substantially alter the estimate of a parameter, or the outcome of a specific test. In contrast, an extreme observation follows the general pattern of the data, but it appears in the extremes of the data set. We name both extreme and outlier observations "atypical" observations. Barnett and Lewis (1994) [9] also present methods to deal with such observations in statistical analysis of data. Some of the proposed procedures exits

Figure 5-4: Explained variation measures as a function of the covariate effect in the model, normally distributed covariate. In the bottom graph, survival times are randomly censored according to an exponential distribution for the censoring times.



to accommodate the atypical observations, while others aim at identifying them.

As Barnett and Lewis (1994) [9] explain, atypical observations do not inevitably "perplex" or "mislead"; they are not necessarily "bad" or "erroneous". Rejection of such observations is a naive way of dealing with them. As Cook and Weisberg (1980) [17] argue the presence of atypical observations does not necessarily imply that they should be deleted or down weighted. This can only be justified when such observations have arisen from purely deterministic reasons such as mistakes in reading or recording in the data (Barnett and Lewis (1994) [9]). Otherwise, they might provide useful information about, first, the underlying mechanism that generated the data and, second, the choice

of the statistical technique or the model applied to analyse the data. For example, if atypical observations proved to be discordant on an assumed normal distribution, it is more likely that we would have chosen to reject them. However, this action might not be justified if an appropriately sophisticated non-normal model incorporated them in a non-discordant fashion. Moreover, in the context of clinical research, they might lead us to unsuspected prognostic factors of particular importance. New procedures have recently been developed in many statistical methods to deal with such observations.

To study the influence of atypical observations on the predictive ability measures, Draper and Smith (1998) [21] explained why outliers reduce the R^2 in linear regression. Wilcox and Muska (1999) [114] presented a non-parametric alternative to the R^2 in linear regression. Rousseeuw and Leroy (1987) [87] presented R^2_R (equation 2.12) for robust regression.

In the context of survival analysis, the issue of atypical observations have not been studied extensively. While there are established methods in normal linear regression models to deal with atypical observations, the methods to diagnose and deal with such observations in the context of survival analysis are not well established. Nardi and Schemper (1999) [72] proposed new residuals which can be used to detect such observations. Pettitt and Bin-Daud (1989) [82] showed that various use of residuals have only limited value in reflecting atypical observations and suggested other forms of diagnostic plots. Henderson and Oman (1993) [44] introduced a method, akin to Cook distance in linear regression, to detect atypical observations in the context of survival analysis. But no method has been universally accepted to detect and accommodate such observations in the context of survival analysis.

The aim of this section is to show the behaviour of the proposed measures when atypical observations are present in the data. To do this, we carried out further studies to investigate the impact of extreme and outlier observations on explained variation measures. The outline of the simulation study and the corresponding results are presented in the following sections.

The study was carried out for four covariate effects, ($\beta = 0.223, 0.405, 0.693, 1.387$), four censoring proportions ($0\%, 20\%, 50\%, 80\%$), and three sample size conditions ($200, 500, 1000$) with $2,000$ replicates in each experimental conditions. In the simulation study, the conditional survival times are generated by assuming $T|X \sim$**Exponential**$(\exp(\beta X))$ where $X \sim N(0,1)$ is the covariate.

To contaminate the data sets with extreme and outlier observations, we used the rule of thumb introduced by Tukey (1977) ([110], page 44). He defined a "mild" outlier observation as an observation that lies 1.5 to 3 times outside the interquantile range, IQR, ($IQR = Q_3 - Q_1$ where Q_1 and Q_3 are the first and third quantile, respectively), and an "extreme" outlier observation as an observation that lies more than 3 times outside the interquantile range, IQR. In the simulation study, we created contaminated data sets containing one atypical observation by replacing one covariate's observation with m times the standard deviation of the covariate, i.e. $m = 1, 2, ..., 8$. Therefore, according to Tukey's definition of "mild" and "extreme" outlier observations, the data sets in which one covariate's observation is replaced with 3 and 4 contain a "mild" outlier observation, whereas the data sets in which one covariate's observation is replaced with values more than 4 contain an "extreme" outlier observation. Finally, we generated the survival times as described in section 4.3.9 depending on what type of atypical observation we study, i.e. extreme or outlier. The survival time of an extreme observation depends on the outlier covariate value. In contrast, the survival time of the outlier observation is independent of the outlier covariate. Random non-informative right censoring with a specified proportion of censored observations was created, as described in section 4.3.4.

We only present the result of one experimental condition where the covariate effect, β, is equal to 0.69, sample size is equal to 200, and censoring proportion is 50%. Similar results were observed in the other experimental conditions. However, the simulation results for the other experimental conditions showed that atypical observations have more impact on the measures in small sample sizes than large ones.

5.7.1 Impact of extreme observations

To generate extreme observations, first the random variable $X \sim N(0, 1)$ was generated. Then, X was contaminated by replacing one observation's covariate with m times the standard deviation of the covariate, i.e. $X \sim N(0, 1)$. Finally, conditional survival times $T|X$, where X is contaminated covariate, were generated based on the procedure described in section 4.3.9. Random non-informative right censoring with a specified proportion of censored observations was created, as described in section 4.3.4. The simulation study carried out for different m values as $m = 1, 2, ...8$.

Figure 5-5 displays the impact of extreme observation on the expected value of explained variation measures by $m*SD$, where SD is the standard deviation of the standard

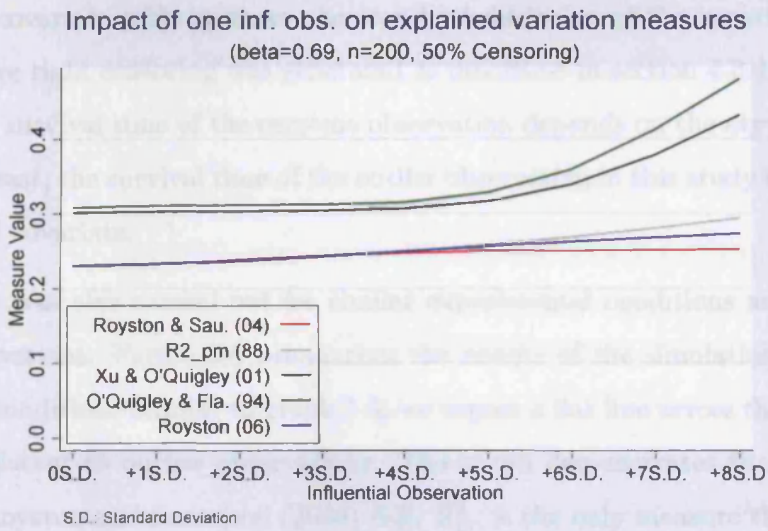


Figure 5-5: Mean of the sampling distribution of explained variation measures as the extreme observation becomes more influential.

normal distribution. For example, $4SD$ in the X axis represents the condition where one observation's covariate, $X \sim N(0,1)$, is replaced with 4, and the corresponding value in the Y axis represents the expected value of measures. The expected value of measures in the uncontaminated data is represented with $0 * SD$ on the X axis.

If a measure is resistant to the extreme observations, its expected value would not change in the presence of such observations. In other words, we expect a flat line across the X axis if the measure is resistant to extreme observations. The graph shows that the measure proposed by Royston and Sauerbrei (2004) [93], R^2_D, is resistant to the extreme observation in the data. The graph shows that the expected value of R^2_D increases slightly but remains constant as the observation becomes more extreme. There is a similar pattern in R^2_{PM} and $R^2_{Royston}$ with more impact on the measure in stronger extreme observations. Furthermore, the measures proposed by O'Quigley and Flandre (1994) [75], R^2_{OQF}, and Xu and O'Quigley (2001) [78], R^2_{XuOQ}, remain constant in small or moderate contamination. But, they increase rapidly when the observation becomes more extreme. Similar results were found in other experimental conditions.

5.7.2 Impact of outlier observations

To generate outlier observations, a normally distributed random variable, i.e. $X \sim N(0,1)$, was generated. Conditional survival times $T|X$ were generated based on the

procedure described in section 4.3.9. Then, the data was contaminated by replacing one observation's covariate with m times the standard deviation of the covariate. Random non-informative right censoring was generated as described in section 4.3.4. In the previous study, the survival time of the extreme observation depends on the atypical covariate value. In contrast, the survival time of the outlier observation in this study is independent of the atypical covariate.

This study was also carried out for similar experimental conditions as the study on extreme observations. Figure 5-6 summarises the results of the simulation study in one experimental condition. Similar to graph 5-5, we expect a flat line across the X axis if the measure is resistant to outlier observations. The graph demonstrates that the measure proposed by Royston and Sauerbrei (2004) [93], R^2_D, is the only measure that is resistant to outliers. The other measures decrease as the outlier contamination becomes stronger.

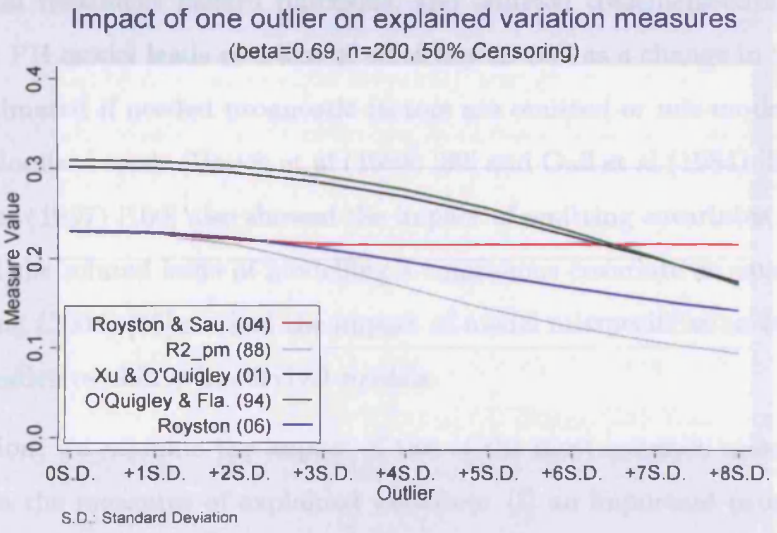


Figure 5-6: Mean of the sampling distribution of explained variation measures as the outlier observation becomes more influential.

The justification for the results of this study is that R^2_{PM} in equation 2.24 depends on the variance of the prognostic index, $Var(\beta' X)$, of the model. Variance is sensitive to extreme and outlier observations; therefore, the presence of such observation has an impact on the R^2_{PM}, whereas R^2_D based on the D measure is not unduly influenced by a small number of atypical observations (Royston and Sauerbrei (2004) [93]) in the data. The D measure is unaffected by outliers, but also by any monotonic increasing transformation of the linear predictor. It is simultaneously a strength and a weakness

of R^2_D. Both R^2_{OQF} and R^2_{XuOQ} depend on the variance of the covariate; therefore, they suffer from a similar problem to that of R^2_{PM}. The measure proposed by Royston (2006) [88], $R^2_{Royston}$, is based on the likelihood function, which is not resistant to atypical observations.

5.8 Impact of model mis-specification on the measures

To analyse the survival of patients in comparative randomised clinical trials, important prognostic factors comprising demographic information such as age, sex, previous medical history, and other medical assessments may be included for the adjustment of the treatment effect. Omitting these factors in the survival model can be considered as one type of model mis-specification. Lagakos and Schoenfeld (1984) [57] described three types of mis-specification of the Cox PH regression models as: omitted or mis-modelled covariates, non-proportional treatment hazard functions, and omitted treatment-covariate interactions. The Cox PH model leads to a loss of efficiency as well as a change in the treatment effect being estimated if needed prognostic factors are omitted or mis-modelled from the analysis of randomised trials (Hauch et al (1998) [39] and Gail et al (1984) [29]). Schmoor and Schumacher (1997) [100] also showed the impact of omitting covariates from the Cox PH model, and the related issue of modelling a continuous covariate as categorical. Rosthoj and Keiding (2004) [86] studied the impact of model misspecification on some of the measures of predictive ability in survival models.

In this section, we examine the impact of two of the most common mis-specifications of the model on the measures of explained variation: (i) an important prognostic factor is omitted from the analysis; and (ii) the true relationship between the prognostic factor and the outcome, log relative hazards in the Cox PH model, is ignored. However, one might argue that these issues are generally dealt with in the model building stage. Nevertheless, understanding the impact of model mis-specification on the measures of explained variation gives better insight into the measures.

5.8.1 Impact of under-fitting - covariate omission

We studied the impact of under-fitting on the measures of explained variation through a series of simulation studies. In the simulation study, we generated pseudo-random, exponentially distributed observations with hazard $\exp(1.386 * X_1 + 0.693 * X_2)$, i.e.

Table 5.11: The expected value of explained variation measures for full and under-fitted models. Normally distributed covariate(s) and random censoring. The figures in brackets are the standard deviation of the sampling distribution.

Measure	Covariates in model	0% Censoring	20% Censoring	50% Censoring
R^2_{PM}	both X_1&X_2	0.594 (0.028)	0.595 (0.029)	0.595 (0.032)
	only X_1 *	0.418 (0.034)	0.432 (0.035)	0.454 (0.039)
R^2_D	both X_1&X_2	0.592 (0.028)	0.593 (0.030)	0.593 (0.034)
	only X_1 *	0.417 (0.034)	0.431 (0.036)	0.453 (0.040)
R^2_{OQF}	both X_1&X_2	0.622 (0.029)	0.636 (0.032)	0.658 (0.036)
	only X_1 *	0.491 (0.034)	0.515 (0.036)	0.553 (0.042)
R^2_{XuOQ}	both X_1&X_2	0.622 (0.029)	0.627 (0.032)	0.640 (0.045)
	only X_1 *	0.491 (0.034)	0.499 (0.039)	0.520 (0.060)
$R^2_{Royston}$	both X_1&X_2	0.534 (0.028)	0.570 (0.032)	0.627 (0.039)
	only X_1 *	0.378 (0.030)	0.416 (0.035)	0.481 (0.045)

*=under-fitted model

$T|X \sim$**Exponential**$(\exp(1.386*X_1+0.693*X_2))$, where $X_i \sim N(0,1)$, $i = 1,2$ are the two independent covariates. The simulations were carried out in three censoring conditions, 0%, 20%, and 50%, with 500 sample size and 2,000 replicates in each experimental condition. Random non-informative right censoring was generated as described in section 4.3.4. Then, the measures are computed for the full model, both X_1 and X_2 in the model, and for the model with only one covariate, X_1. Results from this study are displayed in table 5.11.

The results of simulation studies in section 5.3 indicate that the expected values of R^2_{PM}, R^2_D, and R^2_{XuOQ} are unaffected by the amount of random censoring in the normally distributed covariate. We, however, observe that under-fitted models impose bias on these measures under different censoring conditions, and the bias depends on the amount of censoring. In particular, estimates with the covariates omitted, only X_1 in the model, will be biased toward zero compared to the estimator with covariates included, both X_1&X_2 in the model. This reflects the similar effect of under-fitting on the estimated parameters in the Cox PH model, reported by Gail et al (1984) [29]. It is difficult to quantify the impact of underfitting on R^2_{OQF} and $R^2_{Royston}$ in the censored condition since the simulation studies in section 5.3 showed that even in the full model they increase with the amount of censoring.

One implication of this bias is its impact on the partial measures of predictive ability. The formulae for a partial R^2, similar to the one defined for the linear models, was presented by O'Quigley and Flandre (1994) [75] and O'Quigley et al (2005) [80]. They

introduced the following general formula to compute a partial measure:

$$1 - R^2(X_1, ..., X_p) = \left[1 - R^2(X_1, ..., X_q)\right] \left[1 - R^2(X_{q+1}, ..., X_p | X_1, ..., X_q)\right] \quad (5.1)$$

where $X_1, ..., X_q$ are covariates in the model and $q < p$. In the above equation, the partial measure of explained variation is $R^2(X_{q+1}, ..., X_p | X_1, ..., X_q)$, i.e. the variation in the outcome, survival time, that is explained by the covariates $X_{q+1}, ..., X_p$ after having accounted for the effects of $X_1, ..., X_q$.

In the above study partial R^2s , i.e. $R^2(X_2|X_1)$, can be computed using the formulae $1 - R^2(X_1, X_2) = \left[1 - R^2(X_1)\right] \left[1 - R^2(X_2|X_1)\right]$. We can observe from the results presented in table 5.11 that under-fitting imposes further bias on the measures under different censoring proportions which inevitably affects $R^2(X_2|X_1)$.

5.8.2 Impact of covariate mis-modelling

This section studies the implications of covariate mis-modelling for the explained variation measures, specially in the presence of random censoring. For this purpose, we carried out a set of simulation studies to examine the measures if a covariate is mis-modelled in the Cox PH model. The conditional survival times were generated by assuming two functional forms for the covariate. Figure 5-7 demonstrates the two models and the linear predictor distributions of the corresponding models. In model I, we assumed that the true relationship between the covariate and the log hazard ratio is curvature. In model II, the functional form of the covariate is similar to that of the number of positive lymph nodes in Model III proposed by Sauerbrei and Royston (1999) [94] for breast cancer data discussed in section 2.3. The simulations were carried out in three censoring conditions, 0%, 20%, and 50%, with 500 sample size and 2,000 replicates in each experimental condition. The data generation procedure and models used in the simulation studies are described below for both models.

Model I

In this simulation study, conditional survival times $T|X \sim$**Exponential**$(\exp(f_1(X)))$ are generated as described in section 4.3.7 where

$$f_1(X) = 0.932 * X + 0.156 * X^2 + 0.014 * X^3$$

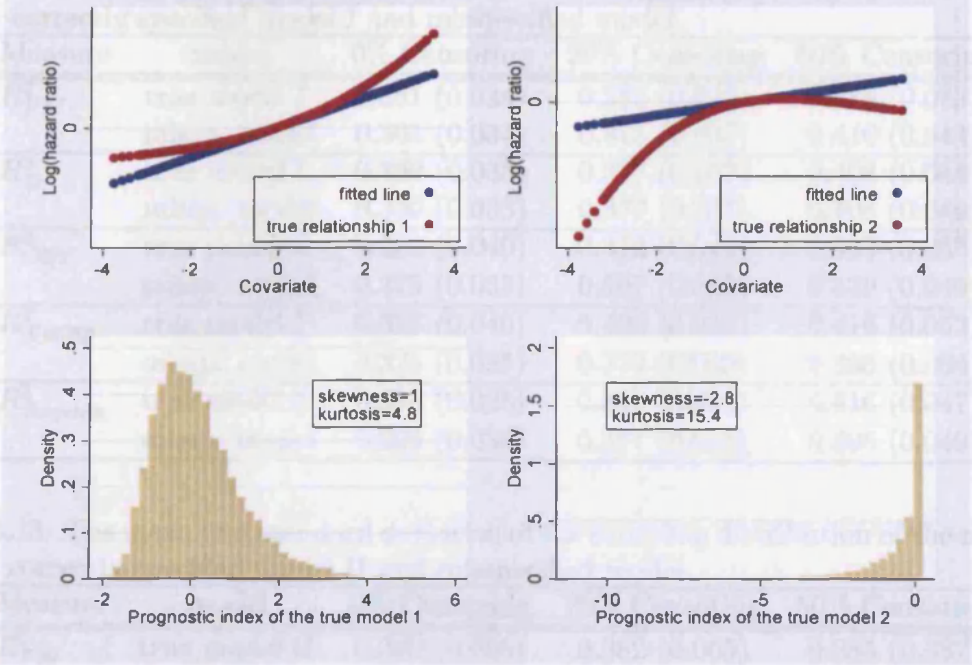


Figure 5-7: The true relationship between the log hazard ratio and the covariate (red curve), and the linear model (blue line) fitted to the simulated data. Bottom graphs show the distribution of prognostic index or linear predictor of the true models.

and $X \sim N(0,1)$. Random non-informative right censoring is generated as described in section 4.3.4. Then, the explained variation measures are computed for 1) the true model and 2) for the mis-specified model where the covariate is modelled as a linear function. The results of simulation study are summarised in table 5.12. The entries in the table are the expected value and standard deviation of the sampling distribution by the amount of censoring.

We can observe that the expected value of R^2_{PM} in the true model is rather consistent across three censoring proportions. In contrast, it increases with censoring in the mis-specified model. As it is apparent from the simulation results, R^2_D is resistant to covariate mis-modelling since the expected value of R^2_D in the true model and the mis-specified model coincide. However, the measure increases with censoring in both true and mis-specified models. As table 5.12 illustrates, this measure is the only one in this category which possesses this property, as long as the relationship between the prognostic index and the log hazard ratio is monotonic (see table 5.13 when this relationship is non-monotonic). The estimates of R^2_{OQF}, R^2_{XuOQ}, and $R^2_{Royston}$ are lower in mis-specified model compared with the corresponding estimates in the true model.

Table 5.12: The mean and standard deviation of the sampling distribution of the measures for the correctly specified model I and misspecified model.

Measure	model	0% Censoring	20% Censoring	50% Censoring
R^2_{PM}	true model I	0.381 (0.034)	0.383 (0.036)	0.388 (0.043)
	missp. model	0.361 (0.034)	0.378 (0.037)	0.410 (0.044)
R^2_D	true model I	0.360 (0.033)	0.377 (0.037)	0.408 (0.044)
	missp. model	0.360 (0.033)	0.377 (0.037)	0.408 (0.044)
R^2_{OQF}	true model I	0.398 (0.040)	0.412 (0.043)	0.443 (0.053)
	missp. model	0.375 (0.035)	0.397 (0.039)	0.439 (0.049)
R^2_{XuOQ}	true model I	0.398 (0.040)	0.402 (0.041)	0.416 (0.053)
	missp. model	0.375 (0.035)	0.379 (0.039)	0.395 (0.064)
$R^2_{Royston}$	true model I	0.314 (0.028)	0.347 (0.034)	0.416 (0.047)
	missp. model	0.295 (0.030)	0.327 (0.035)	0.395 (0.049)

Table 5.13: The mean and standard deviation of the sampling distribution of the measures for the correctly specified model II and misspecified model.

Measure	model	0% Censoring	20% Censoring	50% Censoring
R^2_{PM}	true model II	0.380 (0.055)	0.382 (0.065)	0.385 (0.087)
	missp. model	0.159 (0.033)	0.144 (0.031)	0.126 (0.031)
R^2_D	true model II	0.200 (0.029)	0.181 (0.029)	0.158 (0.031)
	missp. model	0.160 (0.030)	0.145 (0.028)	0.127 (0.029)
R^2_{OQF}	true model II	0.639 (0.068)	0.623 (0.079)	0.589 (0.103)
	missp. model	0.328 (0.038)	0.322 (0.040)	0.299 (0.050)
R^2_{XuOQ}	true model II	0.639 (0.068)	0.636 (0.085)	0.622 (0.132)
	missp. model	0.328 (0.038)	0.328 (0.041)	0.324 (0.073)
$R^2_{Royston}$	true model II	0.312 (0.034)	0.284 (0.035)	0.244 (0.040)
	missp. model	0.192 (0.031)	0.180 (0.032)	0.153 (0.034)

Model II

The simulation structure described above is applied for this study as well, except that the data are generated from a model with linear predictor $f_2(X)$ where

$$f_2(X) = 0.668 * X - 0.413 * X^2 + 0.045 * X^3$$

and $X \sim N(0, 1)$. Table 5.13 demonstrates that the estimates in the true model and mis-specified model differ substantially for all the measures. It is apparent that R^2_{XuOQ} does hardly change with the amount of censoring in both the true and mis-specified models. The estimates of R^2_{PM} also does not change for the true model. But other measures decrease with increasing amount of censoring in both models. Furthermore, unlike model (I) which considered a monotonic function for the covariate, in non-monotonic functions R^2_D results in different values for the true and mis-specified models.

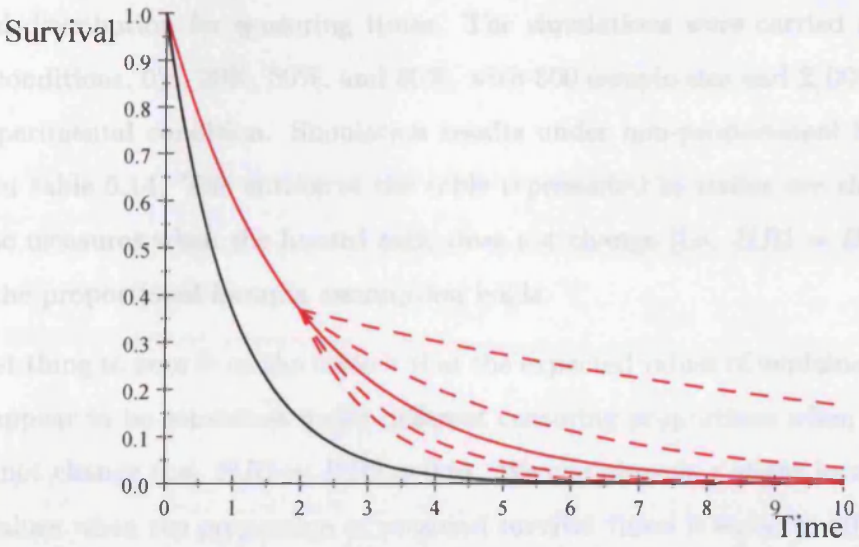


Figure 5-8: The survival pattern of a two-arm trial under non-proportional hazards. Red curve is the survival in the treatment arm, and the black curve is the survival in the control arm. In the treatment arm, the hazard changes for those who survived after two years.

5.8.3 Non-proportional hazards

In this section, a different simulation study was carried out to study the impact of non-proportional hazards on explained variation measures. We used the method proposed by Barthel et al (2006) [10] to generate survival times under non-proportional hazards where the hazards of one arm changes after a specific time in a clinical trails. As they argued, the situations may occur in, for example, a two arm trial when a treatment is very effective in the beginning but patients experience a levelling off of the treatment effect, which in turn brings the survival curves closer together over time or if, such as in a trial comparing surgery followed by chemotherapy with surgery alone, the two treatments have similar hazards in the beginning which then diverge over time (graph 5-8). In this case, the hazard in the treatment arm was changed for each patient who had survived two years in the trial, which led to a change in the overall hazard ratio from $HR1$ to $HR2$. This was simulated by first assigning a probability to whether patients experienced an event before the time of changing hazard. If not, the exponential survival distribution was adapted to incorporate a change in hazards after this point.

Design specifications for all sets of simulations were two years of accrual, two years of follow-up, equal allocation to both treatment arms, exponential survival times, one year

median survival in the control group, and the survival times were censored by assuming exponential distribution for censoring times. The simulations were carried out in four censoring conditions, 0%, 20%, 50%, and 80%, with 500 sample size and 2,000 replicates in each experimental condition. Simulation results under non-proportional hazards are displayed in table 5.14. The entries of the table represented in *italics* are the expected value of the measures when the hazard ratio does not change (i.e. $HR1 = HR2 = 0.5$), i.e. when the proportional hazards assumption holds.

The first thing to note from the table is that the expected values of explained variation measures appear to be consistent under different censoring proportions when the hazard ratio does not change (i.e. $HR1 = HR2 = 0.5$). We can observe a slight increase in the expected values when the proportion of censored survival times is equal to 80%. Second, the value of R^2_D is in line with the values of R^2_{OQF} and R^2_{XuOQ} when the covariate is dichotomy, whereas R^2_{PM} and $R^2_{Royston}$ have smaller values. Third, the impact of non-proportional hazards on the measures diminishes as the amount of censoring increases.

Table 5.14: Simulation results for non-proportional hazards. HR1 - hazard ratio in favour of treatment arm for the first two years in trial. HR2 - hazard ratio after two years in trial. Sample size is 500 in all experimental conditions, and survival times are randomly censored. The standard deviations are given in brackets

Measure	$HR1$	$HR2$	0% Cens.	20% Cens.	50% Cens.	80% Cens.
R^2_{PM}	0.5	0.1	0.149 (0.024)	0.137 (0.026)	0.097 (0.028)	0.075 (0.042)
	0.5	0.3	0.100 (0.019)	0.094 (0.022)	0.082 (0.027)	0.075 (0.042)
	0.5	*0.5*	*0.069 (0.016)*	*0.069 (0.019)*	*0.071 (0.024)*	*0.075 (0.042)*
	0.5	0.7	0.049 (0.014)	0.053 (0.017)	0.063 (0.023)	0.075 (0.041)
	0.5	0.9	0.035 (0.012)	0.041 (0.015)	0.057 (0.022)	0.075 (0.041)
R^2_{D}	0.5	0.1	0.216 (0.024)	0.199 (0.026)	0.144 (0.028)	0.112 (0.042)
	0.5	0.3	0.148 (0.019)	0.14 (0.022)	0.122 (0.027)	0.112 (0.042)
	0.5	*0.5*	*0.104 (0.016)*	*0.105 (0.019)*	*0.107 (0.024)*	*0.112 (0.042)*
	0.5	0.7	0.074 (0.014)	0.081 (0.017)	0.095 (0.023)	0.111 (0.041)
	0.5	0.9	0.054 (0.012)	0.064 (0.015)	0.086 (0.022)	0.111 (0.041)
R^2_{OQF}	0.5	0.1	0.167 (0.028)	0.162 (0.033)	0.138 (0.04)	0.121 (0.066)
	0.5	0.3	0.131 (0.025)	0.129 (0.029)	0.123 (0.039)	0.120 (0.066)
	0.5	*0.5*	*0.101 (0.023)*	*0.104 (0.027)*	*0.111 (0.037)*	*0.120 (0.066)*
	0.5	0.7	0.077 (0.021)	0.084 (0.026)	0.101 (0.036)	0.120 (0.065)
	0.5	0.9	0.057 (0.02)	0.068 (0.024)	0.092 (0.035)	0.120 (0.065)
R^2_{XuOQ}	0.5	0.1	0.167 (0.028)	0.168 (0.032)	0.169 (0.040)	0.120 (0.111)
	0.5	0.3	0.131 (0.025)	0.132 (0.027)	0.135 (0.042)	0.115 (0.111)
	0.5	*0.5*	*0.101 (0.023)*	*0.102 (0.026)*	*0.106 (0.041)*	*0.110 (0.113)*
	0.5	0.7	0.077 (0.021)	0.077 (0.025)	0.084 (0.039)	0.107 (0.111)
	0.5	0.9	0.057 (0.020)	0.058 (0.023)	0.064 (0.037)	0.102 (0.112)
$R^2_{Royston}$	0.5	0.1	0.125 (0.019)	0.129 (0.026)	0.099 (0.030)	0.076 (0.043)
	0.5	0.3	0.090 (0.017)	0.090 (0.021)	0.083 (0.028)	0.076 (0.043)
	0.5	*0.5*	*0.065 (0.015)*	*0.068 (0.018)*	*0.072 (0.025)*	*0.076 (0.043)*
	0.5	0.7	0.047 (0.014)	0.052 (0.017)	0.064 (0.024)	0.076 (0.042)
	0.5	0.9	0.035 (0.012)	0.041 (0.015)	0.057 (0.023)	0.076 (0.042)

5.9 Discussion

In this chapter, we studied the measures of explained variation through a set of simulation studies. The simulations have been aimed at finding how these measures perform in different conditions, addressing the unresolved issues with respect to properties presented in tables 3.1 to 3.2 of chapter 3. Furthermore, we studied the impact of model misspecification that might occur in statistical analysis of survival data.

We first evaluated the measures in non-censored data to have an understanding of the proposed measures in terms of effect size and spread of the sampling distribution. This has revealed the impact of the covariate distribution on these measures. Table 5.1 showed that R^2_{PM} is the only measure that was independent of covariate distribution or prognostic index in the Cox PH model. The measure proposed by Royston and Sauerbrei (2004), R^2_D, depends on covariate distribution; the measure results in lower values if the covariate distribution departs from normality. The more the departure from normality, the lower the expected value of the measure. The measure decreases about $35\% - 45\%$ in the skewed covariate distributions considered in this study depending on the covariate effect. This reflects the properties of D measure [93] in a model with a non-normal covariate distribution.

Both R^2_{OQF} and R^2_{XuOQ} depend on covariate distribution; the measures result in lower values if covariate distribution is positively skewed and higher values if covariate distribution is negatively skewed. For example, in non-censored data when $\beta = 1.386$, the expected values of both measures is 0.597 in covariates with positively skewed distributions, whereas they increase to 0.728 in negatively skewed distributions. $R^2_{Royston}$ also changes as the covariate distribution alters with no specific pattern evident from the simulation studies. However, the change in the expected value of this measure in different covariate distributions is not as much as that of R^2_D, R^2_{OQF}, and R^2_{XuOQ}.

The impact of censoring was investigated by considering different censoring mechanisms and censoring proportions. Table 5.15 summarises the findings of simulation studies presented in section 5.3. For the majority of the measures, the impact of censoring depends on the skewness of covariate distribution in the model. The codes in the table show the extent of censoring effect on the measures, with 1 representing almost no effect, i.e. the average percentage change in the expected value of the measure is $0\% - 9\%$ compared to that of non-censored data, and 4 representing a large effect, i.e. the average percentage change in the expected value of the measure is over 50% compared to that of

Table 5.15: Summary of censoring effects on explained variation measures by the covariate distribution and censoring type. The codes show the extent of censoring effect in different situations from almost no effect, 1, to a large effect, 4.

Censoring type	Exp. Var. measure	Covariate or Prognostic Index Distribution			
		Normal	Lognormal	Pos. skewed	Neg. skewed
Random censoring	R^2_{PM}	1	1	1	1
	R^2_{D}	1	2	3	2
	R^2_{OQF}	2	2	2	2
	R^2_{XuOQ}	1	1	1	1
	$R^2_{Royston}$	2	3	4	2
Type I censoring	R^2_{PM}	1	1	1	1
	R^2_{D}	1	2	4	3
	R^2_{OQF}	2	2	2	3
	R^2_{XuOQ}	2	2	2	3
	$R^2_{Royston}$	2	3	4	3

1: Almost no effect, i.e. the average percentage change in the mean of sampling distribution is 0%−9%

2: Slight effect, i.e. the average percentage change in the mean of sampling distribution is 10%−19%

3: Moderate effect, i.e. the average percentage change in the mean of sampling distribution is 20%−49%

4: Large effect, i.e. the average percentage change in the mean of sampling distribution is over 50%

non-censored data. The table indicates that R^2_{PM} is the only measure which is independent of censoring in all covariate distributions. It also shows that $R^2_{Royston}$ is the poorest measure with respect to the impact of censoring.

The sampling distribution of the measures were displayed for different covariate effects under different censoring and sample size conditions in section 5.4. Histograms of the sampling distribution of the measures indicate that the measure proposed by Royston and Sauerbrei (2004) [93], R^2_D, can be regarded as a consistent estimator because its sampling distribution becomes more concentrated around the expected value as the sample size becomes larger in all covariate effects and censoring proportions (figure 5-3) . The figures of other measures in this category show similar findings, with the exception of R^2_{XuOQ} which results in negative values as censoring increases.

Sample size seems to affect the measures by only a modest amount if the effective sample size, i.e. the number of events, k, is small, i.e. $k \simeq 40$ in our simulation studies. If the covariate effect is small, i.e. $\beta = 0.223$, all of the measures increase by about 25% when the total sample size and number of events are 200 and 40, respectively.

All of the measures possess the parameter monotonicity property which requires the measures to increase as the covariate effect becomes stronger. Although the expected values of the measures do not decrease by adding a new covariate to the model, table

Measure	Sample Size	Does parameter monotonicity hold?
R^2_{PM}	no effect[1]	yes[2]
R^2_D	no effect[1]	yes[2]
R^2_{OQF}	no effect[1]	yes[2]
R^2_{XuOQ}	no effect[1]	yes[2]
$R^2_{Royston}$	no effect[1]	yes[2]

1) There is a moderate effect of sample size on measures only when covariate effect is 1.25, sample size is 200, and censoring proportion is high, i.e. 80%.

2) The measure increases with increasing parameter effect.

Table 5.16: Summary of sample size effect and parameter monotonicity property of explained variation measures.

5.10 shows that the $R^2_{Royston}$ is the only measure that is strictly monotonic, i.e. it always satisfies the number of variables monotonicity. In all the replicates, $R^2_{Royston}$ does not decrease as a new covariate is added to the model. Among the measures, R^2_{XuOQ} performs the poorest in this regard since the proportion of simulations in which this measure decreased after adding one or two independent covariates to the model was the highest compared with other measures.

The simulations to study the impact of extreme and outlier observations revealed that R^2_D is the only measure which remains almost unaffected by such observations. Other measures generally increase in the presence of extreme observations, whereas they decreased in the presence of outlier observations in the data. The results of our simulation studies indicate that in the presence of severe outlier observations, i.e. $m = 8$ in section 5.7, R^2_{PM}, R^2_D, R^2_{OQF}, R^2_{XuOQ}, and $R^2_{Royston}$ decrease by about 59%, 6%, 44%, 44%, and 38% respectively (the expected values of measures at $8SD$ are compared with the corresponding values at $0SD$, no contamination, in figure 5-6), whereas they increase by 27%, 8%, 55%, 43%, and 18% in the presence of influential extreme observations (figure 5-5).

All measures attain values near 1 for large, but plausible, values of covariate effects, i.e. βs. The simulation study presented in section 5.6 shows that the measures are an increasing function of β when the covariate or PI of the model is normally distributed. The rate of increase slows down after $\beta = 3.47$.

Finally, we studied the impact of model mis-specification on the measures in section 5.8. As described by Lagakos and Schoenfeld (1984) [57], model mis-specification of the Cox PH regression models includes non-proportional treatment hazard functions, omitted and mis-modelled covariates. Section 5.8.1 demonstrates that omission of influential

covariates in a model imposes bias on the measures. Furthermore, covariate mis-modelling affects the measures, depending on how severe the departure is from the true functional form. The simulation studies showed that R^2_D is the only measure that results in the same value for both the true model and a mis-specified one if the true relationship between the covariate and the outcome is monotonic. The impact of non-proportional hazards was discussed in section 5.8.3. Table 5.14 demonstrates that all measures are susceptible to changes in treatment hazards. The susceptibility of the measures to non-proportional hazards diminishes as the amount of censoring increases.

In summary, our study showed that R^2_{PM} is independent of censoring and covariate distribution, but it is very sensitive to covariate outliers in the data. R^2_D performs well generally, but struggles with heavily skewed covariate(s) when the amount of censoring is high, i.e. more than 50%. R^2_{OQF} performs reasonably well in general, but it is not a consistent estimator of the population value, $\mathbf{R}^2_{OQF}$ as expressed in equation 2.29, in the presence of censoring. To overcome this, R^2_{OQF} was further developed to introduce R^2_{XuOQ}. However, R^2_{XuOQ} possesses the undesirable property of resulting in negative values as censoring increases. Finally, $R^2_{Royston}$ has the poorest performance with regard to the essential properties outlined in chapter 3 compared with other measures in this category. Therefore, the two explained variation measures R^2_{PM} and R^2_D can be recommended for general use, depending on the skewness of the covariate, or prognostic index, of the model and the amount of censoring. R^2_{PM} is suitable if the amount of censored observations in the data is high, i.e. about $70\% - 90\%$, and the covariate, or prognostic index, of the model is heavily skewed. The measure proposed by Royston and Sauerbrei (2004), R^2_D, is preferable if there is an indication of extreme covariate outliers in the data.

The next chapter studies the measures of explained randomness in a similar fashion.

Chapter 6

Investigation of the measures of explained randomness

6.1 Introduction

This chapter studies various aspects of potentially recommendable measures in the explained randomness category. The measures in this category generally quantify the randomness or uncertainty in the outcome, as defined in equation 2.30 of chapter 2, that is explained by prognostic factors in a regression model. The measures are ρ^2_W, ρ^2_{XuOQ}, and ρ^2_k proposed by Kent and O'Quigley (1988) [49], Xu and O'Quigley (1999) [116], and O'Quigley et al (2005) [80], respectively. Since no explicit formula is available for ρ^2_W, Kent and O'Quigley (1988) [49] suggested $\rho^2_{W,A}$, in equation 2.36, as an approximation . They, however, did not compare $\rho^2_{W,A}$ and ρ^2_W in detail. We include $\rho^2_{W,A}$ in our studies in sections 6.2 and 6.3 to elucidate its performance and compare it to the other explained randomness measures. $\rho^2_{W,A}$ is not intrinsically an explained randomness measure and, in principle, is similar to R^2_{PM}. The only difference between them is that the variance of error term in the definition of R^2_{PM} in equation 2.24, i.e. $\frac{\pi^2}{6} \simeq 1.645$, is replaced with 1 in the definition of $\rho^2_{W,A}$ in equation 2.36.

This chapter has a similar structure to that of chapter 5. We carried out the same simulation studies on the above explained randomness measures, hence the study design of the simulations for each section is the same as those presented in chapter 5. We, therefore, present the results through similar graphs and tables to describe the main findings for each measure, and do not explain the simulation study design for each study

again.

In summary, this chapter addresses the following:

- The expected value of the measures in non-censored data
- The impact of different covariate distributions on the measures
- The impact of censoring on the measures
- Consistency, distributional shape, and sample size effect
- Monotonicity properties of the measures
- The upper bound of the measures
- The impact of extreme and outlier observations on the measures
- The impact of model mis-specification on the measures

In addition, we evaluated Kullback-Leibler information gain for the Cox PH model, and hence developed a new measure of explained randomness for the proportional hazards models. Since the main theme of this thesis is to compare the already proposed measures of predictive ability in survival models, we only present the new measure in Appendix B.8 of this thesis. The last section contains the discussion of this chapter's findings.

6.2 Impact of covariate distribution on the measures

In this section, we present the results of our simulation study, carried out to assess the measures in the absence of censoring. The study was also carried out to investigate the expected value and dispersion of measures in different covariate effects and covariate distributions. In summary, the simulations were run for four covariate distributions, four covariate effects $\beta = \{0.223, 0.405,\ 0.693, 1.386\}$ representing hazard ratios of $\{1.25, 1.5, 2, 4\}$, and three sample size conditions $n = \{200, 500, 1000\}$, with $5,000$ replicates in each experimental condition.

The results of the simulations are summarised in tables 6.1 to 6.3. The first thing to note from table 6.1 is that the explained randomness measures generally result in higher values compared with the corresponding values of the explained variation measures, presented in table 5.1. As it is evident from tables 6.1 to 6.3, the measures lead to

similar results in the normally distributed covariate. In this case, the expected value and the relative spread of the sampling distribution, as expressed in terms of the *C.V.*, of the measures is similar. The results of the simulation study show that $\rho^2_{W,A}$ is a good approximation to ρ^2_W if the covariate is normally distributed, otherwise they differ. O'Quigley et al (2005) [80] showed that in the absence of censoring ρ^2_k and ρ^2_{XuOQ} coincide, and in censored data ρ^2_k can be considered as a good approximation to ρ^2_{XuOQ}. Tables 6.1 to 6.3 display this theory.

Tables 6.2 and 6.3 show that the spread of the sampling distribution of ρ^2_W, ρ^2_{XuOQ}, and ρ^2_k are similar in normally, lognormally, and positively skewed distributions. In negatively skewed distributions, however, the spread of the sampling distribution of ρ^2_W and $\rho^2_{W,A}$ is higher than that of ρ^2_k and ρ^2_{XuOQ}. Some important findings for each measure is explained in the following sections.

Table 6.1: Mean of the sampling distribution of explained randomness measures by the covariate distribution and covariate effect across all sample size conditions, censoring=0%

Covariate Distribution	$\exp(\beta)$	ρ^2_W	$\rho^2_{W,A}$	ρ^2_{XuOQ}	ρ^2_k
normal	1.25	0.049	0.050	0.048	0.048
	1.5	0.141	0.143	0.134	0.134
	2	0.316	0.325	0.296	0.296
	4	0.637	0.657	0.602	0.602
lognormal	1.25	0.046	0.050	0.045	0.045
	1.5	0.128	0.143	0.125	0.125
	2	0.282	0.325	0.275	0.275
	4	0.584	0.656	0.575	0.574
positively skewed	1.25	0.042	0.051	0.042	0.042
	1.5	0.110	0.144	0.109	0.109
	2	0.235	0.324	0.233	0.233
	4	0.495	0.652	0.486	0.485
negatively skewed	1.25	0.062	0.049	0.052	0.052
	1.5	0.195	0.142	0.142	0.142
	2	0.433	0.322	0.292	0.292
	4	0.759	0.651	0.552	0.551

6.2.1 Kent and O'Quigley measures (1988) - ρ^2_W & $\rho^2_{W,A}$

This measure varies from 0.049 to 0.637 in the normally distributed covariate. It decreases in positively skewed distributions, whereas it increases in negatively skewed distributions. In contrast, its proposed approximation, $\rho^2_{W.A}$, is not affected by the changes in the covariate distribution. As it is evident from table 6.3, the spread of the sampling distribution of this measure and its approximation, $\rho^2_{W,A}$, decreases as the covariate effect becomes

Table 6.2: Standard deviation of the sampling distribution of explained randomness measures by the covariate distribution and covariate effect across all sample size conditions, censoring=0%

Covariate Distribution	$\exp(\beta)$	ρ^2_W	$\rho^2_{W,A}$	ρ^2_{XuOQ}	ρ^2_k
normal	1.25	0.022	0.022	0.021	0.021
	1.5	0.034	0.035	0.032	0.032
	2	0.041	0.042	0.039	0.039
	4	0.032	0.031	0.032	0.032
lognormal	1.25	0.020	0.023	0.020	0.020
	1.5	0.030	0.035	0.030	0.030
	2	0.036	0.044	0.037	0.037
	4	0.031	0.034	0.034	0.033
positively skewed	1.25	0.018	0.024	0.018	0.018
	1.5	0.027	0.040	0.028	0.028
	2	0.036	0.056	0.037	0.037
	4	0.042	0.051	0.047	0.040
negatively skewed	1.25	0.038	0.025	0.025	0.025
	1.5	0.079	0.044	0.039	0.039
	2	0.105	0.063	0.046	0.046
	4	0.070	0.056	0.040	0.040

larger.

6.2.2 Xu and O'Quigley measure (1999) - ρ^2_{XuOQ}

Similar findings are observed for this measure, with the exception that this measure leads to slightly lower values than ρ^2_W in all the covariate distributions.

6.2.3 O'Quigley et al measure (2005) - ρ^2_k

O'Quigley et al (2005) [80] showed that this measure converges to the same values as ρ^2_{XuOQ} in non-censored data. The summary data presented in tables 6.1 to 6.3 confirms this theory. Therefore similar conclusions to those of ρ^2_{XuOQ} can be drawn for this measure in non-censored data.

6.3 Impact of censoring on the measures

In this section, we study the impact of censoring on the explained randomness measures through a series of simulation studies similar to those used to assess the impact of censoring on the explained variation measures in section 5.3. In summary, the simulations were run for two types of censoring mechanisms, type I and random censoring,

Table 6.3: Coefficient of variation of explained randomness measures by the covariate distribution and covariate effect, expressed as %. Table entries are the average across all combinations of sample sizes, censoring=0%.

Covariate Distribution	$\exp(\beta)$	ρ^2_W	$\rho^2_{W,A}$	ρ^2_{XuOQ}	ρ^2_k
normal	1.25	42.1	42.3	41.7	41.7
	1.5	22.8	23.0	22.7	22.7
	2	12.2	12.3	12.4	12.4
	4	4.7	4.5	5.0	5.0
lognormal	1.25	41.1	42.7	41.8	41.8
	1.5	22.3	23.4	22.6	22.6
	2	12.2	12.8	12.8	12.8
	4	5.1	5.0	5.5	5.5
positively skewed	1.25	41.1	44.9	41.4	41.4
	1.5	23.6	26.7	24.1	24.1
	2	14.7	16.4	15.0	15.0
	4	8.0	7.4	9.5	7.7
negatively skewed	1.25	59.0	47.7	46.1	46.1
	1.5	39.9	29.7	25.9	25.9
	2	23.5	18.6	14.9	14.8
	4	8.8	8.1	6.9	6.8

and four censoring proportions, 0%, 20%, 50%, and 80%, four covariate effects $\beta = \{0.223, 0.405, 0.693, 1.386\}$ representing hazard ratios of $\{1.25, 1.5, 2, 4\}$, and three sample size conditions $n = \{200, 500, 1000\}$, with 5,000 replicates in each experimental condition. The mechanisms applied for generating each censoring type were explained in section 4.3.4 of chapter 4.

Tables 6.4 to 6.6 summarise the results of the simulation study on the proposed explained randomness measures. The entries in tables 6.4 and 6.5 are the average over two censoring types, four covariate effects, and three sample size conditions. The values in table 6.6 are the average across four censoring proportions, four covariate effects, and three sample size conditions. The figures in these tables are the average across four covariate effects, and three sample size conditions. In summary, it is evident that ρ^2_W and $\rho^2_{W,A}$ are least affected by censoring, whereas ρ^2_k is most affected by the amount of censoring in all covariate distributions.

Detailed simulation results are presented in Appendix A. The tables in Appendix A show the impact of censoring by the covariate distribution, censoring type, and censoring proportion in a similar way to table 6.6. The impact of censoring on each measure is explained in details in the following sections.

Table 6.4: The average percentage difference from the expected value of explained randomness measures in the corresponding non-censored data by the covariate distribution and censoring proportion.

Covariate Distribution	% Censored	ρ^2_W	$\rho^2_{W,A}$	ρ^2_{XuOQ}	ρ^2_k
normal	20	0.3	0.3	2.7	4.5
	50	1.3	1.3	6.0	10.9
	80	5.2	5.3	13.5	21.7
lognormal	20	0.2	0.2	5.4	8.9
	50	0.8	0.9	14.1	23.7
	80	3.0	3.3	31.8	47.2
positively skewed	20	0.1	0.1	8.3	14.3
	50	0.3	0.4	25.3	43.0
	80	1.5	1.8	62.9	93.6
negatively skewed	20	1.3	1.2	-7.4	-9.8
	50	4.1	3.8	-15.7	-19.5
	80	13.9	13.6	-22.6	-23.0

Table 6.5: Coefficient of variation of explained randomness measures by the covariate distribution and censoring proportion, expressed as %.

Covariate Distribution	% Censored	ρ^2_W	$\rho^2_{W,A}$	ρ^2_{XuOQ}	ρ^2_k
normal	20	22.2	22.3	22.9	23.3
	50	27.3	27.4	29.1	30.0
	80	41.4	41.7	47.3	48.1
lognormal	20	21.4	22.2	23.2	24.0
	50	25.1	26.0	30.0	32.3
	80	35.5	36.7	50.7	53.8
positively skewed	20	22.5	24.5	25.4	26.5
	50	24.7	26.8	33.1	37.5
	80	31.4	33.7	59.2	65.9
negatively skewed	20	36.7	30.0	24.4	23.4
	50	44.9	38.8	29.2	26.6
	80	69.2	64.9	42.6	39.1

6.3.1 Kent and O'Quigley measures (1988) - ρ^2_W & $\rho^2_{W,A}$

Censoring has almost no effect on this measure, except in highly censored data, i.e. 80% censoring, with negatively skewed covariates. The average percentage change in the expected value of the measure is generally less than 5% compared with the expected value of the measure in the corresponding non-censored data. We observe a slight increase in the expected value of the measure in positively skewed covariates with highly censored data, i.e. 80% censoring; the measure is on average 13.9% higher compared with the corresponding non-censored data. Table 6.5 shows that the spread of the sampling distribution increases with the amount of censoring, as expected. Random and type I (administrative) censoring have similar impact on this measure (table 6.6). The results

Table 6.6: Summary performance of explained randomness measures by the covariate distribution and censoring mechanism.

Measure	Covariate Distribution	Random Censoring		Type I Censoring	
		Average % Difference	*C.V.*	Average % Difference	*C.V.*
ρ^2_W	normal	2.4	30.6	2.2	30.0
	lognormal	1.5	27.8	1.1	26.9
	positively skewed	0.9	26.7	0.3	25.7
	negatively skewed	6.1	49.8	6.7	50.7
$\rho^2_{W,A}$	normal	2.4	30.8	2.2	30.2
	lognormal	1.7	28.7	1.3	27.8
	positively skewed	1.1	28.9	0.5	27.8
	negatively skewed	5.8	44.0	6.5	45.2
ρ^2_{XuOQ}	normal	1.1	32.3	13.7	34.0
	lognormal	4.8	32.0	29.4	37.2
	positively skewed	9.6	34.1	54.8	44.4
	negatively skewed	-9.9	35.1	-19.6	29.1
ρ^2_k	normal	11.0	33.6	13.7	34.0
	lognormal	23.8	36.2	29.4	37.2
	positively skewed	45.6	42.5	54.9	44.1
	negatively skewed	-15.4	30.4	-19.5	29.1

of our simulation study show similar impact of censoring on the approximation of this measure, $\rho^2_{W.A}$.

6.3.2 Xu and O'Quigley measure (1999) - ρ^2_{XuOQ}

Table 6.4 displays that, overall, ρ^2_{XuOQ} is affected by the amount of censoring. However, table 6.6 reveals that the effect is mainly as a result of type I censoring, and random censoring has almost no effect on this measure since the average percentage difference in the expected value of the measure is less than 10% compared to that of non-censored data. Therefore ρ^2_{XuOQ} performs well in random censoring. In the type I censoring, the measure increases rapidly with the amount of censoring in positively skewed covariates, whereas it decreases rapidly in negatively skewed covariates. The higher the amount of censoring, the larger the impact on the measure. Table 6.5 reveals that the spread of the sampling distribution also increases with the amount of censoring.

6.3.3 O'Quigley et al measure (2005) - ρ^2_k

This measure is affected by both random and type I censoring. It appears that there is an interaction between censoring and the covariate distribution in this measure. The

measure increases with the amount of censoring in positively skewed covariates, whereas decreases in negatively skewed covariates. Table 6.6 shows that type I censoring has similar impact on this measure to that of ρ^2_{XuOQ}.

6.4 Consistency, distributional shape, and sample size effect

In this section, we investigate the consistency and the shape of the sampling distribution of the measures of explained randomness, together with the impact of sample size. The consistency of the proposed measures, ρ^2_W, ρ^2_{XuOQ}, and ρ^2_k, are discussed by Kent & O'Quigley (1988) [49], Xu & O'Quigley (1999) [116], and O'Quigley et al (2005) [80]. We first summarise their findings on the consistency of the proposed measures. Then, we illustrate the shape of the sampling distribution of the measures in the presence of censoring. Finally, the effect of sample size on the measure is studied. The approximation of ρ^2_W, $\rho^2_{W,A}$, possesses similar properties to that of R^2_{PM} and will not be discussed in this section.

6.4.1 Consistency of the measures

For models beyond the normal linear regression, Kent (1983) [50] proposed a general measure of correlation, known as explained randomness measure ρ^2_{IG}, based on the transformation of the Kullback-Leibler information gain [55]

$$\rho^2_{IG} = 1 - \exp\{-\Gamma(\beta)\}$$

where $\Gamma(\beta)$ is twice the Kullback-Leibler information gain [55], as described in section 2.3.2. Therefore, all the proposed explained randomness measures are a transformation of the Kullback-Leibler information gain [55]. The only difference between them is the way the Kullback-Leibler information gain, $\Gamma(\beta)$, is defined for each measure. The measures are considered consistent if the estimator of their corresponding Kullback-Leibler information gain, $\widehat{\Gamma}(\beta)$ is consistent.

For the measure proposed by Kent and O'Quigley (1988) [49], ρ^2_W, the Kullback-Leibler information gain is defined as

$$\Gamma_1(\beta) = 2\int_X \int_T \log\left\{\frac{f^*(t|X;\beta)}{f^*(t|X;0)}\right\} f^*(t|X;\beta)dtdF(x)$$

where $f^*(t|X;\beta) = \alpha\exp(\mu+\beta X)t^{\alpha-1}\exp\left[-t^\alpha\exp\left(\mu+\beta X\right)\right]$ and $F(x)$ is the distribution function of the covariate X. Kent and O'Quigley (1988) [49] showed that $\Gamma_1(\beta)$ can be consistently estimated by the fitted information gain

$$\widehat{\Gamma}_1(\widehat{\beta}) = \frac{2}{n}\sum_{i=1}^{n}\int_T \log\left\{\frac{f^*(t|x_i;\widehat{\beta})}{f^*(t|x_i;0)}\right\} f^*(t|X;\widehat{\beta})dt$$

where $\widehat{\beta}$ is the maximum likelihood estimator of β. Thus, $\widehat{\rho}^2_W = 1-\exp\{-\widehat{\Gamma}_1(\widehat{\beta})\}$ is a consistent estimator of the population value ρ^2_W.

For the measure proposed by Xu and O'Quigley (1999) [116], ρ^2_{XuOQ}, the Kullback-Leibler information gain is defined as

$$\Gamma_2(\beta) = 2\int_T \int_X \log\left\{\frac{g(x|t;\beta)}{g(x|t;0)}\right\} g(x|t;\beta)dxdF(t)$$

where $F(t)$ is the marginal distribution function of T, and $g(x|t;.)$ is the conditional density or conditional probability function of the covariate, X, given T. Xu and O'Quigley (1999) [116] showed that $\Gamma_2(\beta)$ can be consistently estimated by

$$\widehat{\Gamma}_2(\widehat{\beta}) = 2\sum\nolimits_{j=1}^{k} W(t_j)\sum\nolimits_{i=1}^{n}\pi_i(t_i;\widehat{\beta})\log\left\{\frac{\pi_i(t_i;\widehat{\beta})}{\pi_i(t;0))}\right\}$$

where $W(t_j)$ the jump in the Kaplan-Meier curve at event time t_j and

$$\pi_j(t;\beta) = \frac{Y_j(t)\exp(\beta Z_j)}{\sum_{l=1}^{n} Y_l(t)\exp(\beta Z_l)}.$$

In the measure proposed by O'Quigley et al (2005) [80], ρ^2_k, $\Gamma_2(\beta)$ is also defined as the Kullback-Leibler information gain. However, O'Quigley et al (2005) [80] proposed an alternative estimator for $\Gamma_2(\beta)$ as

$$\widehat{\Gamma}_{2A}(\widehat{\beta}) = \frac{2}{k}\sum_{i=1}^{n}\delta_i\log\left\{\frac{\pi_i(X_i;\widehat{\beta})}{\pi_i(X_i;0)}\right\}$$

where $\delta_i = I(T_i < C_i)$ and k is the number of events. O'Quigley et al (2005) [80] showed that in the absence of censoring $\widehat{\Gamma}_2(\widehat{\beta})$ and $\widehat{\Gamma}_{2A}(\widehat{\beta})$ will converge to the same population

values, hence both estimators are consistent estimator of $\Gamma_2(\beta)$. They, however, converge to different quantities in the presence of censoring, but can be anticipated to be close (O'Quigley et al (2005) [80]).

O'Quigley and Flandre (2006) [76] claimed that the advantage of ρ^2_{XuOQ}, which is based upon $\widehat{\Gamma}_2(\widehat{\beta})$, is that it is consistent in both censored and non-censored data and does not depend upon censoring. Although ρ^2_k is not consistent in the presence of censoring, it is particularly straightforward to evaluate, being a simple transformation of the partial likelihood ratio statistic. The results of our simulation studies, summarised in table 6.6, indicate that ρ^2_{XuOQ} is not affected by censoring if the survival times are randomly censored, but it results in the same values as ρ^2_k in the presence of type I or administrative censoring.

6.4.2 Sampling distribution of the measures

Figure 6-1 illustrates the sampling distribution of Kent and O'Quigley measure (1988), ρ^2_W, from our simulation study, by the covariate effect, sample size, and censoring proportion, with $5,000$ replicates in each experimental condition. The covariate is normally distributed and the survival times are randomly censored by considering an exponential distribution for censoring times, as described in section 4.3.4.

As in explained variation measures, we crudely explore the sampling distributions of the estimators graphically over the range of n in this study. Sampling distribution of consistent estimators should tend towards a spike over the parameter of interest as n becomes ever larger. All distributions in figure 6-1 appear to exhibit this tendency. For example, for the normally distributed covariate when the covariate effect, β, is 1.39, the expected value of the measure proposed by Kent and O'Quigley (1988), ρ^2_W, is 0.637 (table 6.1). Figure 6-1 demonstrates that the distribution of ρ^2_W is approximately centred over the expected value. The distribution of the measure is clearly becoming more concentrated and spiking near this value as sample size increases. The shape of the distribution of the measures proposed by Xu and O'Quigley (1999) and O'Quigley et al (2005) [80], ρ^2_{XuOQ} and ρ^2_k follows a similar pattern, i.e. they display considerable skewness when censoring is more than 50%.

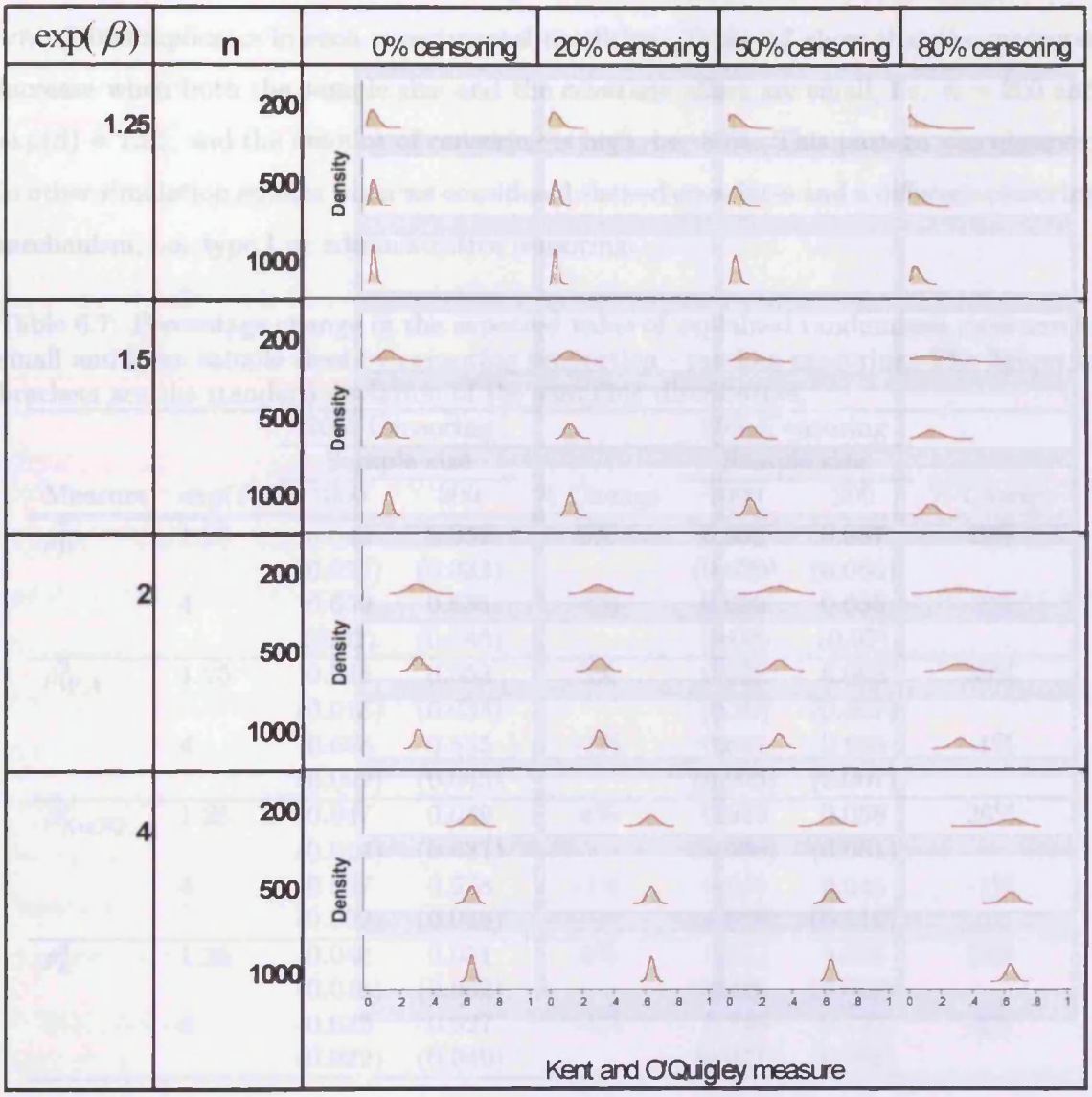


Figure 6-1: Sampling distributions of Kent & O'Quigley measure (1988) by the covariate effect, sample size, and censoring proportions in the normally distributed covariate and random censoring condition.

6.4.3 Impact of sample size on the measures

Similar to the effect of sample size on explained variation measures, the simulation results show that the measures of explained randomness increase slightly when the effective sample size, i.e. number of events, is small. We tabulated the results of the simulations in table 6.7 where the covariate is normally distributed and the data is randomly censored, with 5,000 replicates in each experimental condition. Table 6.7 show that the measures increase when both the sample size and the covariate effect are small, i.e. $n = 200$ and $\exp(\beta) = 1.25$, and the amount of censoring is high, i.e. 80%. This pattern was observed in other simulation studies when we considered skewed covariates and a different censoring mechanism, i.e. type I or administrative censoring.

Table 6.7: Percentage change in the expected value of explained randomness measures in small and large sample sizes by censoring proportion - random censoring. The figures in brackets are the standard deviation of the sampling distribution.

		20% Censoring			80% Censoring		
		Sample size			Sample size		
Measure	$\exp(\beta)$	1000	200	% Change	1000	200	% Change
ρ^2_W	1.25	0.048	0.052	8%	0.052	0.067	29%
		(0.015)	(0.033)		(0.029)	(0.066)	
	4	0.639	0.635	-1%	0.639	0.635	-1%
		(0.02)	(0.046)		(0.03)	(0.07)	
$\rho^2_{W,A}$	1.25	0.048	0.053	8%	0.052	0.068	29%
		(0.015)	(0.034)		(0.29)	(0.067)	
	4	0.658	0.655	-1%	0.658	0.655	-1%
		(0.002)	(0.045)		(0.003)	(0.007)	
ρ^2_{XuOQ}	1.25	0.047	0.049	4%	0.046	0.058	26%
		(0.014)	(0.031)		(0.028)	(0.061)	
	4	0.607	0.598	-1%	0.649	0.643	-1%
		(0.022)	(0.048)		(0.078)	(0.119)	
ρ^2_k	1.25	0.048	0.051	6%	0.053	0.068	28%
		(0.015)	(0.032)		(0.03)	(0.069)	
	4	0.635	0.627	-1%	0.746	0.733	-2%
		(0.022)	(0.049)		(0.037)	(0.086)	

6.5 Monotonicity properties of the proposed measures

This section consists of two parts discussing the two monotonicity properties in explained randomness measures, as defined in chapter 3. In the second part where we discuss the number of variables monotonicity, similar simulation study to that of explained variation measures is carried out. The sample size is 500, and 2,000 replicates are generated

for each experimental condition. In the simulation, the distribution of survival time is generated using the algorithm outlined in section 4.3.9 by assuming only one covariate that is normally distributed.

6.5.1 Parameter monotonicity

The parameter monotonicity property of ρ^2_W and ρ^2_{XuOQ} was analytically established by Kent & O'Quigley (1988) [49] and Xu and O'Quigley (1999) [116]. The measure proposed by O'Quigley et al (2005) [80], ρ^2_k, is similar to ρ^2_{XuOQ} in non-censored data [80], hence satisfying parameter monotonicity property. Furthermore, the simulation results presented in table 6.1 and graph 6-2 demonstrate that the measures increase as the covariate effect becomes stronger.

6.5.2 Number of variables monotonicity

The number of variables monotonicity means that the expected value of a suitable measure of predictive ability should not decrease by adding new covariates to the model. Tables 6.8 and 6.9 demonstrate the results of similar simulation study as in section 5.5 to investigate the number of variables monotonicity of the measures. The following models are fitted after generating the data: Model I with one dependent covariate; Model II with only dependent covariate and one independent covariate; and Model III with only dependent covariate and two independent covariates. The entries in table 6.8 are the differences in the expected values of the measures after fitting models II and III compared to model I. The table shows that the expectation of the measures does not decrease after adding new covariates to the model.

Table 6.9 displays the proportion of simulations in which the measures decreased after adding one and two independent covariates by covariate effects and censoring proportions. Whilst ρ^2_k always increases after adding a new covariate to the model in censored and non-censored data, ρ^2_{XuOQ} does not always increase in censored data.

6.6 Upper bound of the measures

In this section, we demonstrate the upper bound of the measures of explained randomness using similar simulation studies to those of section 5.6. Figure 6-2 contains two graphs

Table 6.8: Mean difference in the expected value of the measures after adding one or two independent covariates to the model in 2,000 simulations, normally distributed covariates.

Measure	exp(β)	Model II		Model III	
		Mean difference to model I		Mean difference to model I	
		0% censoring	80% censoring	0% censoring	80% censoring
ρ^2_W	1.25	0.002	0.009	0.004	0.018
	1.5	0.002	0.008	0.004	0.015
	2	0.002	0.005	0.003	0.011
	4	0.001	0.002	0.002	0.005
$\rho^2_{W,A}$	1.25	0.002	0.009	0.004	0.018
	1.5	0.002	0.008	0.004	0.016
	2	0.002	0.006	0.003	0.011
	4	0.001	0.003	0.002	0.005
ρ^2_{XuOQ}	1.25	0.002	0.009	0.004	0.019
	1.5	0.002	0.008	0.003	0.017
	2	0.001	0.007	0.003	0.013
	4	0.001	0.003	0.002	0.005
ρ^2_k	1.25	0.002	0.008	0.004	0.016
	1.5	0.002	0.007	0.003	0.014
	2	0.001	0.005	0.003	0.010
	4	0.001	0.002	0.002	0.004

which summarise the results of simulation studies in both non-censored and censored data. In the simulation studies, survival times are exponentially distributed, the covariate is normally distributed $X \sim N(0,1)$, sample size is 500, and non-informative random censoring is generated by considering an exponential distribution for the censoring times with $2,000$ replicates in each experimental condition. Figure 6-2 displays the expected value of the measures from $\beta = 0.22$ ($\exp(\beta) = 1.25$) to $\beta = 5.55$ ($\exp(\beta) = 256$) for 0% and 50% censoring. it is evident that the expected value of the measures increases as the covariate effect becomes larger, and they reach values close to 1 for high but reasonable covariate effects.

6.7 Robustness of the measures

In this section, simulation studies analogous to section 5.7 are carried out to investigate the impact of "atypical" observations, i.e. extreme and outlier observations as described in section 5.7, on the explained randomness measures. Similarly, this section consists of two parts which demonstrate the impact of extreme and outlier observations on the measures of explained randomness respectively. The methods we apply to contaminate the data with extreme and outlier observations are described in section 5.7. Similarly, we present the results of the simulation studies through graphs. In the graphs, the X

Table 6.9: Proportion decrease in measures after adding one or two independent covariates to the model in 2000 simulations, normally distributed covariates.

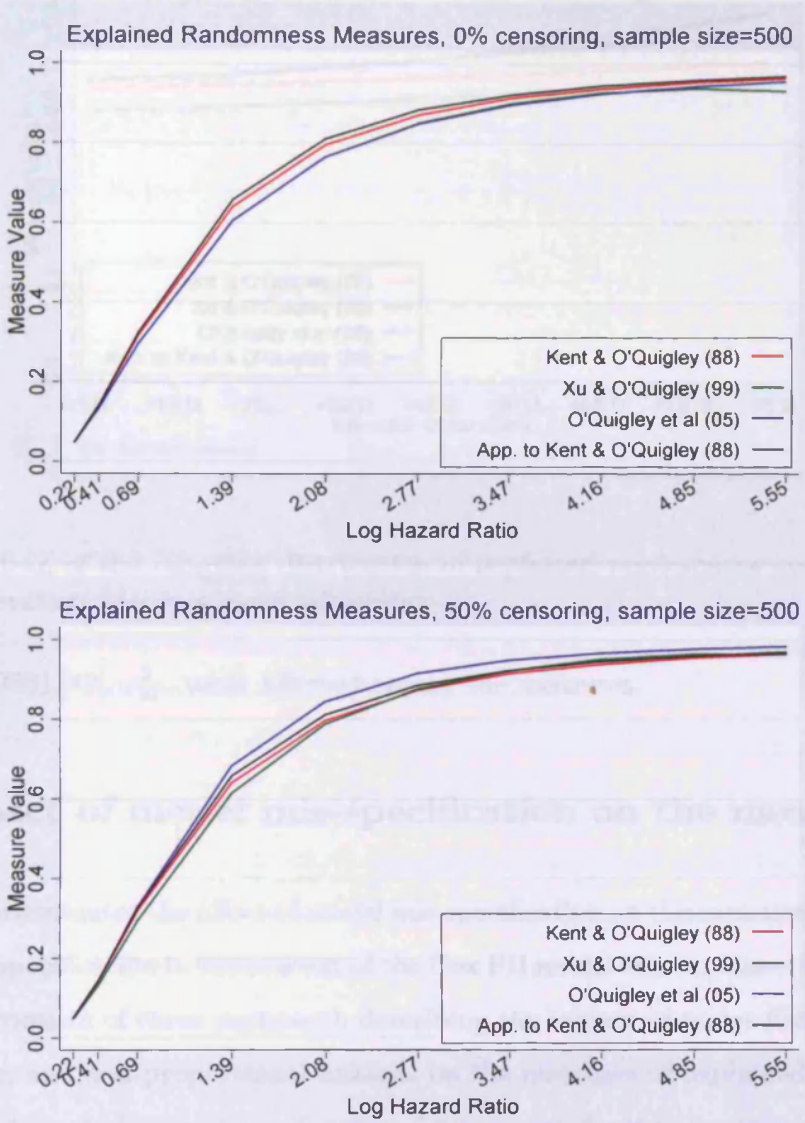
Measure	exp(β)	Model II		Model III	
		prop. decreased to model I		prop. decreased to model I	
		0% censoring	80% censoring	0% censoring	80% censoring
ρ^2_W	1.25	0.11	0.07	0.03	0.02
	1.5	0.15	0.13	0.07	0.05
	2	0.20	0.21	0.11	0.12
	4	0.23	0.29	0.13	0.21
$\rho^2_{W,A}$	1.25	0.10	0.07	0.03	0.02
	1.5	0.15	0.12	0.06	0.04
	2	0.18	0.19	0.09	0.10
	4	0.20	0.27	0.09	0.18
ρ^2_{XuOQ}	1.25	0.00	0.12	0.00	0.05
	1.5	0.00	0.21	0.00	0.11
	2	0.00	0.27	0.00	0.19
	4	0.00	0.35	0.00	0.28
ρ^2_k	1.25	0.00	0.00	0.00	0.00
	1.5	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00

axis represents the extent of contamination. For example, $5SD$ in the X axis represents the condition where one observation's covariate, $X \sim N(0,1)$, is replaced with 5, and the corresponding value in the Y axis represents the expected value of the measures in this condition. We present the result of one experimental condition where the covariate effect, β, is equal to 0.69, sample size is equal to 200, and censoring proportion is 50%. We also carried out simulation studies considering different sample sizes, i.e. 500 and 1000, and censoring proportions, i.e. 0%, 20% and 80%. Similar results were observed in other experimental conditions. However, the results suggested that both extreme and outliers observations have more impact on the measures in small sample sizes than large ones, as expected.

6.7.1 Impact of extreme observations

Figure 6-3 displays the impact of one extreme observation on the expected value of explained randomness measures. If the measures are resistant to the extreme observations, the expected value of the measures would not change in the presence of such observations. In other words, we expect a flat line across the X axis if the measure is resistant to extreme observations. The graph demonstrates that the measures are not resistant to extreme observations, i.e. the covariate and corresponding time move towards the

Figure 6-2: Explained randomness measures as a function of the covariate effect in the model, normally distributed covariate. In the bottom graph, survival times are randomly censored according to an exponential distribution for censoring times.



extremes of their respective distributions. The graph indicates that the measures go up as one observations becomes more extreme.

6.7.2 Impact of outlier observations

A similar simulation study was also carried out to demonstrate the impact of outlier observations on ρ^2_W, ρ^2_{XuOQ}, and ρ^2_k. In the simulation studies the data sets are generated using the method explained in section 5.7.2. Figure 6-4 demonstrates that the measures are largely influenced by such observations in the data. The measures decrease as the outlier contamination becomes more severe, with the measure proposed by Kent and

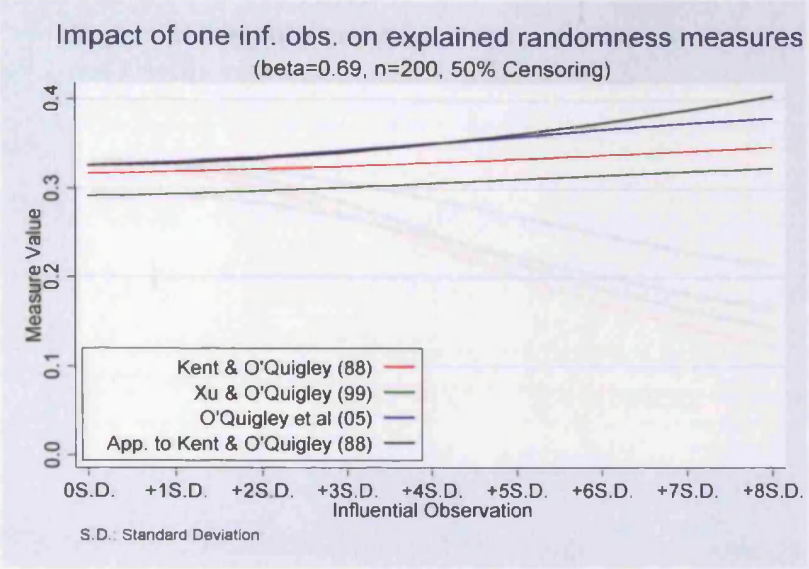


Figure 6-3: Mean of the sampling distribution of explained randomness measures as the extreme observation becomes more influential.

O'Quigley (1988) [49], ρ^2_W, most affected among the measures.

6.8 Impact of model mis-specification on the measures

This section investigates the effect of model mis-specification on the measures. The notion of model mis-specification in the context of the Cox PH model was explained in section 5.8. This section consists of three parts each describing the impact of under-fitting, covariate mis-modelling, and non-proportional hazards on the measures of explained randomness. Similar simulation studies to those of section 5.8 are carried out to investigate the issue of model mis-specification on the measures; therefore, we do not describe the study design in this section again. Likewise, all the simulations are carried out in different censoring conditions with 500 sample size and 2,000 replicates in each experimental condition. The results are summarised in similar tables to those of section 5.8.

6.8.1 Impact of under-fitting - covariate omission

Table 6.10 demonstrates the impact of under-fitting on the explained randomness measures. The entries of the table are the expected value and the standard deviation of the sampling distribution of the measures for full and under-fitted models. We observe that under-fitting imposes bias on ρ^2_W, $\rho^2_{W,A}$, and ρ^2_{XuOQ} in censored data, and the bias in-

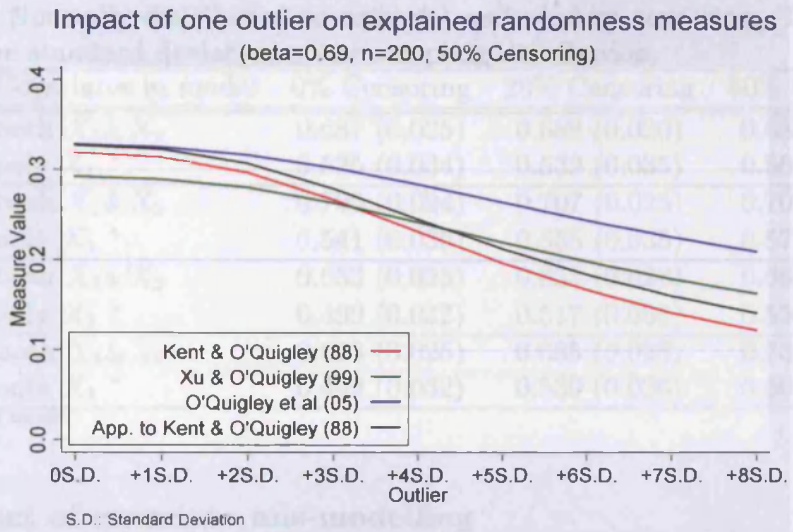


Figure 6-4: Mean of the sampling distribution of the explained randomness measures as the outlier observation becomes more influential.

creases as the proportion of censored observations increases. For example, ρ^2_W and ρ^2_{XuOQ} are fairly consistent in the full model, whereas they are inconsistent in the under-fitted model, i.e. they increase with the amount of censoring.

Similar to the measures of explained variation, the implication of this bias is that it imposes bias on the estimates of the partial measure of explained randomness suggested by Kent and O'Quigley (1988) [49] and O'Quigley et al (2005) [80]. Similar to equation 5.1, they suggested the following general formula to compute the partial measure of explained randomness

$$1 - \rho^2(X_1, ..., X_p) = \left[1 - \rho^2(X_1, ..., X_q)\right] \left[1 - \rho^2(X_{q+1}, ..., X_p | X_1, ..., X_q)\right] \qquad (6.1)$$

where $X_1, ..., X_q$ are covariates in the model and $q < p$. In the above equation, the partial measure of explained randomness is $\rho^2(X_{q+1}, ..., X_p | X_1, ..., X_q)$, i.e. the randomness in the outcome, survival time, that is explained by the covariates $X_{q+1}, ..., X_p$ after having accounted for the effects of $X_1, ..., X_q$.

In the above study, a partial measure of explained randomness $\rho^2(X_2|X_1)$ can be computed using the formula $1 - \rho^2(X_1, X_2) = \left[1 - \rho^2(X_1)\right] \left[1 - \rho^2(X_2|X_1)\right]$. The results presented in table 6.10 indicate that under-fitting imposes further bias on the measures under different censoring proportions which inevitably affects $\rho^2(X_2|X_1)$.

Table 6.10: The expected value of explained randomness measures for full and under-fitted models. Normally distributed covariate(s) and random censoring. The figures in brackets are the standard deviation of the sampling distribution.

Measure	Covariates in model	0% Censoring	20% Censoring	50% Censoring
ρ^2_W	both X_1&X_2	0.687 (0.025)	0.688 (0.026)	0.688 (0.029)
	only X_1 *	0.525 (0.034)	0.539 (0.035)	0.560 (0.038)
$\rho^2_{W,A}$	both X_1&X_2	0.706 (0.024)	0.707 (0.025)	0.707 (0.028)
	only X_1 *	0.541 (0.035)	0.555 (0.035)	0.577 (0.038)
ρ^2_{XuOQ}	both X_1&X_2	0.653 (0.025)	0.655 (0.028)	0.668 (0.045)
	only X_1 *	0.499 (0.032)	0.517 (0.035)	0.550 (0.050)
ρ^2_k	both X_1&X_2	0.653 (0.025)	0.685 (0.028)	0.733 (0.033)
	only X_1 *	0.499 (0.032)	0.539 (0.036)	0.603 (0.043)

*=under-fitted model

6.8.2 Impact of covariate mis-modelling

In a similar study to that of section 5.8.2, we investigate the impact of covariate mis-modelling on the explained randomness measures, i.e. modelling the covariate, X, as linear function of log hazard ratio in the Cox PH model where the true functional form of the covariate is either "model I", $f_1(X)$, or "model II", $f_2(X)$. Figure 5-7 in chapter 5 demonstrates the functional forms of the covariate against the log hazard ratio in the Cox PH model. The findings of simulation studies are summarised below for each model.

Model I

In this model, the true functional form of the covariate in the Cox PH model is:

$$f_1(X) = 0.932 * X + 0.156 * X^2 + 0.014 * X^3$$

where $X \sim N(0, 1)$. Table 6.11 displays the mean and standard deviation of the sampling distribution of the measures for true and mis-specified models by censoring proportions. This table indicates that the measure proposed by Kent and O'Quigley (1988), ρ^2_W, and its approximation, $\rho^2_{W,A}$, result in different values in both the true and mis-specified models. It also suggests that they are fairly consistent under different censoring proportions in the true model. In the mis-specified model, however, they increase as the amount of censoring increases. With increasing censoring, both ρ^2_{XuOQ} and ρ^2_k increase in both true and mis-specified models.

Table 6.11: The mean and standard deviation of the sampling distribution of the measures for correctly specified model I and misspecified model.

Measure	model	0% Censoring	20% Censoring	50% Censoring
ρ^2_W	true model I	0.438 (0.033)	0.441 (0.038)	0.451 (0.057)
	missp. model	0.466 (0.035)	0.484 (0.038)	0.516 (0.044)
$\rho^2_{W,A}$	true model I	0.503 (0.036)	0.504 (0.038)	0.509 (0.046)
	missp. model	0.481 (0.037)	0.499 (0.039)	0.532 (0.045)
ρ^2_{XuOQ}	true model I	0.429 (0.032)	0.431 (0.035)	0.442 (0.050)
	missp. model	0.407 (0.034)	0.425 (0.038)	0.461 (0.052)
ρ^2_k	true model I	0.429 (0.032)	0.466 (0.037)	0.538 (0.048)
	missp. model	0.407 (0.034)	0.444 (0.040)	0.517 (0.052)

Table 6.12: The mean and standard deviation of the sampling distribution of measures for correctly specified model II and misspecified model.

Measure	model	0% Censoring	20% Censoring	50% Censoring
ρ^2_W	true model II	0.630 (0.079)	0.630 (0.091)	0.625 (0.116)
	missp. model	0.231 (0.044)	0.211 (0.042)	0.187 (0.042)
$\rho^2_{W,A}$	true model II	0.501 (0.059)	0.502 (0.068)	0.503 (0.091)
	missp. model	0.236 (0.045)	0.216 (0.042)	0.190 (0.043)
ρ^2_{XuOQ}	true model II	0.427 (0.039)	0.414 (0.044)	0.384 (0.061)
	missp. model	0.280 (0.040)	0.255 (0.038)	0.221 (0.044)
ρ^2_k	true model II	0.427 (0.039)	0.394 (0.042)	0.345 (0.049)
	missp. model	0.280 (0.040)	0.264 (0.042)	0.229 (0.047)

Model II

In this model, the true functional form of the covariate in the Cox PH model is:

$$f_2(X) = 0.668 * X - 0.413 * X^2 + 0.045 * X^3$$

where $X \sim N(0,1)$. Similarly, table 6.12 contains the mean and standard deviation of the sampling distribution of measures for true and mis-specified models by censoring proportions. We observe that in the true model, both ρ^2_W and $\rho^2_{W,A}$ are consistent under different censoring proportions, but they decrease in the mis-specified model as the amount of censoring increases. The simulation results also show that both ρ^2_{XuOQ} and ρ^2_k decrease in true and mis-specified models as the amount of censoring increases.

6.8.3 Non-proportional hazards

In an analogous study to that of explained variation measures in section 5.8.3, we carried out simulation studies to investigate the impact of non-proportional hazards on the explained randomness measures. Design specifications for all sets of simulations were

two years of accrual, two years of follow-up, equal allocation to both treatment arms, exponential survival times, one year median survival in the control group, and the survival times were censored by assuming exponential distribution for censoring times. The simulations were carried out in four censoring conditions, 0%, 20%, 50%, and 80%, with 500 sample size and 2,000 replicates in each experimental condition.

Simulation results under non-proportional hazards are displayed in table 6.13. The entries of the table represented in *italics* are the expected value of the measures when the hazard ratio does not change (i.e. $HR1 = HR2 = 0.5$), i.e. when the proportional hazards assumption holds.

Table 6.13 shows that the measures result in similar values in the Cox PH model with a dichotomous covariate if the proportional hazards assumption holds. Furthermore, the expected value of ρ^2_W and ρ^2_{XuOQ}, presented in *italics,* appear to be consistent under different censoring proportions when the hazard ratio does not change (i.e. $HR1 = HR2 = 0.5$). In this case, the expected value of ρ^2_k increases slightly with the amount of censoring. Finally, the impact of non-proportional hazards on the measures diminishes as the amount of censoring increases.

Table 6.13: Simulation results for non-proportional hazards. HR1 - hazard ratio in favour of treatment arm for the first two years in trial. HR2 - hazard ratio after two years in trial. Sample size is 500 in all experimental conditions, and survival times are randomly censored. The standard deviations are given in brackets

Measure	$HR1$	$HR2$	0% Cens.	20% Cens.	50% Cens.	80% Cens.
ρ^2_W	0.5	0.1	0.212 (0.030)	0.195 (0.032)	0.144 (0.037)	0.111 (0.054)
	0.5	0.3	0.149 (0.027)	0.140 (0.029)	0.121 (0.035)	0.110 (0.054)
	0.5	*0.5*	*0.106 (0.025)*	*0.106 (0.027)*	*0.106 (0.034)*	*0.110 (0.054)*
	0.5	0.7	0.077 (0.022)	0.082 (0.025)	0.095 (0.033)	0.110 (0.054)
	0.5	0.9	0.056 (0.020)	0.065 (0.023)	0.086 (0.031)	0.109 (0.054)
$\rho^2_{W,A}$	0.5	0.1	0.223 (0.032)	0.204 (0.034)	0.149 (0.039)	0.114 (0.057)
	0.5	0.3	0.154 (0.029)	0.145 (0.031)	0.125 (0.037)	0.114 (0.057)
	0.5	*0.5*	*0.109 (0.026)*	*0.109 (0.029)*	*0.110 (0.036)*	*0.114 (0.057)*
	0.5	0.7	0.078 (0.023)	0.085 (0.026)	0.097 (0.034)	0.113 (0.057)
	0.5	0.9	0.057 (0.021)	0.067 (0.024)	0.088 (0.033)	0.113 (0.057)
ρ^2_{XuOQ}	0.5	0.1	0.188 (0.027)	0.174 (0.029)	0.139 (0.039)	0.109 (0.058)
	0.5	0.3	0.140 (0.025)	0.131 (0.028)	0.116 (0.035)	0.107 (0.058)
	0.5	*0.5*	*0.103 (0.024)*	*0.103 (0.026)*	*0.103 (0.034)*	*0.104 (0.057)*
	0.5	0.7	0.076 (0.022)	0.082 (0.025)	0.093 (0.033)	0.102 (0.055)
	0.5	0.9	0.056 (0.020)	0.066 (0.023)	0.085 (0.032)	0.101 (0.055)
ρ^2_k	0.5	0.1	0.188 (0.027)	0.194 (0.033)	0.152 (0.041)	0.116 (0.059)
	0.5	0.3	0.140 (0.025)	0.140 (0.030)	0.127 (0.039)	0.115 (0.059)
	0.5	*0.5*	0.103 (0.024)	0.107 (0.028)	0.111 (0.037)	0.115 (0.058)
	0.5	0.7	0.076 (0.022)	0.084 (0.026)	0.099 (0.035)	0.115 (0.058)
	0.5	0.9	0.056 (0.020)	0.067 (0.024)	0.089 (0.034)	0.114 (0.058)

6.9 Discussion

In this chapter, we studied the measures of explained randomness proposed by Kent and O'Quigley (1988) [49], ρ^2_W, Xu and O'Quigley (1999) [116], ρ^2_{XuOQ}, and O'Quigley et al (2005) [80], ρ^2_k. We repeated similar studies to those performed on the measures of explained variation, presented in chapter 5. This helped us to understand the behaviour of explained randomness measures in similar conditions and to compare the two categories consistently.

The results of simulation studies showed that explained randomness measures studied in this chapter generally result in higher values than the explained variation measures presented in chapter 5. The measures are influenced by the distribution of covariates in the model. They generally lead to higher values in negatively skewed covariates and lower values in positively skewed covariates. Contrary to the claim by Kent and O'Quigley [49], table 6.1 shows that $\rho^2_{W,A}$ is not a good approximation for ρ^2_W if the covariate distribution is asymmetric. We also observed that the measures were in agreement if the covariate distribution is normally distributed.

The simulation results presented in section 6.3 demonstrate that ρ^2_W and its approximation, $\rho^2_{W,A}$, are least affected and ρ^2_k is most affected by the amount of censoring. Also, the impact of censoring on the measures depends on the distribution of covariate as seen in table 6.14. The codes in the table show the extent of the censoring effect on the measures of explained randomness, with 1 representing almost no effect and 4 representing a large effect. The table indicates that ρ^2_W is the only measure which is independent of censoring in all covariate distributions, whereas ρ^2_k is most affected.

The distributional properties of explained randomness measures were investigated in section 6.4.2. The sampling distribution of Kent and O'Quigley's measure (1988) [49], ρ^2_W, is presented in graph 6-1, for different covariate effects and censoring proportions. The sampling distribution of the estimator of ρ^2_W display considerable skewness when censoring is more than 50%. This graph confirms Kent and O'Quigley's theory [49] that this measure is a consistent estimator; the sampling distribution of the estimator becomes more concentrated around the expected value of the measure as the sample size increases. The shape of the sampling distribution of other measures follows a similar pattern.

Our simulation studies indicate that ρ^2_W is the only measure which is consistent under both random and administrative censoring. Sample size has a moderate effect on

Table 6.14: Summary of censoring effects on explained randomness measures by the covariate distribution and censoring type. The codes show the extent of censoring effect in different situations from almost no effect, 1, to a large effect, 4.

Censoring type	Exp. Rand. measure	Covariate or Prognostic Index Distribution			
		Normal	Lognormal	Pos. skewed	Neg. skewed
Random censoring	ρ^2_W	1	1	1	1
	$\rho^2_{W,A}$	1	1	1	1
	ρ^2_{XuOQ}	1	1	2	2
	ρ^2_k	2	3	4	2
Type I censoring	ρ^2_W	1	1	1	1
	$\rho^2_{W,A}$	1	1	1	1
	ρ^2_{XuOQ}	2	3	4	3
	ρ^2_k	2	3	4	2

1: Almost no effect, i.e. the average percentage change in the mean of sampling distribution is 0%−9%

2: Slight effect, i.e. the average percentage change in the mean of sampling distribution is 10%−19%

3: Moderate effect, i.e. the average percentage change in the mean of sampling distribution is 20%−49%

4: Large effect, i.e. the average percentage change in the mean of sampling distribution is over 50%

Table 6.15: Summary of sample size effect and parameter monotonicity of the explained randomness measures.

Measure	Sample Size	Does parameter monotonicity hold?
ρ^2_W	no effect[1]	yes[2]
ρ^2_{XuOQ}	no effect[1]	yes[2]
ρ^2_k	no effect[1]	yes[2]

1) There is a moderate effect of sample size on measures only when covariate effect is 1.25, sample size is 200, and censoring proportion is high, i.e. 80%.

2) The measure increases with increasing parameter effect.

the measures whereby they increase by about 25% if both the effective sample size, i.e. number of events, and the covariate effect are small ($\beta = 0.22$ in table 6.7).

Graphs presented in section 6.6 illustrate that all measures increase as the covariate effect increases, hence satisfying the parameter monotonicity property. Moreover, the results of another simulation study presented in table 6.9 indicate that ρ^2_k is the only measure that is strictly monotonic. Although ρ^2_{XuOQ} possesses the same property in non-censored data, the simulation study showed that when the censoring is 20% and $HR = 4$, the measure decreases in 35% of replicates as a new covariate is added to the model.

The investigation which was carried out in section 6.7, to elucidate the behaviour of the explained randomness measures in the presence of extreme and outlier observations, show that the measures are susceptible to such observations in the data. The measures increase in the presence of extreme observations, whereas they decrease in the presence of outlier observations. However, the results of simulation studies show that the impact

of extreme observations on the explained randomness measures is not as large as those of explained variation measures. For example, in the presence of severe outlier observations, i.e. $m = 8$ in section 6.7, ρ^2_W, ρ^2_{XuOQ}, and ρ^2_k decrease by 59%, 44%, and 36% respectively (the expected value of the measures at $8SD$ are compared with the corresponding values at $0SD$, no contamination, in figure 6-4), whereas they increase by only 10%, 10%, and 15% in the presence of influential extreme observations (figure 6-3).

The graphs presented in section 6.6 indicate that the measures of explained randomness can reach values near 1 in both non-censored and censored data. The measures increase as the covariate effect in the models becomes larger.

The impact of three different types of model mis-specification was investigated in section 6.8. First, the results of the study on the impact of under-fitting, or omitted covariates, show that under-fitting imposes positive bias on the measures in the presence of censoring; the measures increase with an increasing amount of censoring. Second, the simulation study shows that the measures are influenced by covariate mis-modelling, depending how severe the departure is from the true functional form of the covariate. Furthermore, table 6.13 demonstrates that all measures are susceptible to changes in treatment hazards. Similar to the impact of non-proportional hazards on the explained variation measures, the susceptibility of the measures to non-proportional hazards diminishes as the amount of censoring increases. Among the three types of model mis-specification, under-fitting is the most common in practice, which has implications for the partial measure of explained randomness suggested by Kent and O'Quigley (1988) [49] and O'Quigley et al (2005) [80] - it imposes bias on the proposed partial measure of explained randomness in the presence of censoring.

In summary, among the explained randomness measures, the measure proposed by Kent and O'Quigley (1988) [49], ρ^2_W, performs reasonably well with regard to the essential properties outlined in chapter 3. Its approximation, $\rho^2_{W,A}$, performs well with respect to the essential properties but is not a good approximation for ρ^2_W if the covariate distribution is asymmetric. The measure proposed by Xu and O'Quigley (1999) [116], ρ^2_{XuOQ}, performs well with random censoring, but struggles in type I or administrative censoring. The results of our study indicate that among the explained randomness measures, the measure proposed by O'Quigley et al (2005) [80], ρ^2_k, performs worst with regard to our essential properties.

The next chapter presents similar studies on the proposed predictive accuracy mea-

sures and the measure proposed by Schemper and Kaider (1997).

Chapter 7

Investigation of the measures of predictive accuracy

7.1 Introduction

This chapter studies various aspects of potentially recommendable measures in the predictive accuracy category. The measures in this category quantify the ability of the regression model to predict the outcome, i.e. being "alive" or "dead" in the context of survival analysis. The two measures are proposed by Graf et al (1999) [31], $R^2_G(T^*)$, and Schemper and Henderson (2000) [97], V_{SchH}, in this chapter.

We also included the results of similar investigations carried out to evaluate the measure proposed by Schemper and Kaider (1997) [98], R^2_{SchK}. Since this measure it is not based on either a variation function or the Kullback-Leibler information gain [55], it can not be classified as an explained variation or explained randomness measure; neither is it a predictive accuracy measure because it does not evaluate the accuracy of the model-based survival probability predictions. It, however, uses a non-parametric measure of correlation such as Spearman correlation coefficient (Spearman (1904) [108]) or Kendall τ (Kendall (1938) [47]) to provide a measure of association between the imputed survival times and covariates in the model.

The studies performed on the explained variation and explained randomness measures are repeated for the above measures in this chapter, hence the study design in all simulation studies are similar to those of chapters 5 and 6. We will not explain the study design and only present the results through similar graphs and tables.

Like the two previous chapters, this chapter addresses the following:

- The expected value of the measures in non-censored data
- The impact of different covariate distributions on the measures
- The impact of censoring on the measures
- Consistency, distributional shape, and sample size effect
- Monotonicity properties of the measures
- The impact of atypical observations on the measures
- The upper bound of the measures
- The impact of model mis-specification on the measures

As was discussed in section 2.3.3, Graf et al's measure (1999) [31], $R^2_G(T^*)$, evaluates the predictive accuracy of the model at a particular time point, T^*. In practice, the choice of the time point depends on the aim of the study. For example, the aim of the study might be to evaluate the performance of the fitted survival model in predicting the individual's status as "dead" or "alive" after $T^* = 2$ years. In the simulation studies, however, we considered different time points to elucidate the behaviour of this measure at different times. The time points are the $0.10th$, $0.15th$, $0.20th$, $0.25th$, and $0.50th$ quantile of the exponential distribution used to generate survival times, as described in section 4.3.7. This corresponds to 5 time points as $T_1 = 5.27$, $T_2 = 8.13$, $T_3 = 11.16$, $T_4 = 14.38$, $T_5 = 34.66$.

The measures proposed by Schemper and Henderson (2000) [97], V_{SchH}, and Schemper and Kaider (1997) [98], R^2_{SchK}, provide an overall measure of predictive ability.

7.2 Impact of covariate distribution on the measures

Simulations were carried out to assess the measures of predictive accuracy with non-censored data. The study was conducted with the same experimental conditions as those of section 5.2, and the results are presented through similar tables. The simulations were run for four covariate distributions, four covariate effects $\beta = \{0.223, 0.405, 0.693, 1.386\}$ representing hazard ratios of $\{1.25, 1.5, 2, 4\}$, and three sample size conditions, $\{200, 500, 1000\}$, with $5,000$ replicates for each experimental condition.

Tables 7.1 to 7.3 summarise the simulation results for different covariate distributions and covariate effects in non-censored data. They show the expected value, the standard deviation of the sampling distribution, and the relative dispersion of the measures. The first thing to note from the table is that the predictive accuracy measures, $R^2_G(T^*)$ and V_{SchH}, appear to be lower than the corresponding values of the explained variation and explained randomness measures presented in tables 5.1 and 6.1. The only measure in the "other" category proposed by Schemper and Kaider (1997) [98], R^2_{SchK}, seems to be in agreement with explained variation measures in the normally distributed covariate. Some of the findings are summarised in the following sections for each measure.

Table 7.1: Mean of the sampling distribution of predictive accuracy measures and Schemper and Kaider's measure (1997) by the covariate distribution and covariate effect across all sample size conditions, censoring=0%

Covariate		Graf et al measure at different time points						
distribution	$\exp(\beta)$	$R^2_G(T_1)$	$R^2_G(T_2)$	$R^2_G(T_3)$	$R^2_G(T_4)$	$R^2_G(T_5)$	V_{SchH}	R^2_{SchK}
normal	1.25	0.006	0.008	0.011	0.013	0.024	0.027	0.036
	1.5	0.019	0.028	0.035	0.043	0.073	0.069	0.104
	2	0.064	0.087	0.106	0.123	0.176	0.159	0.244
	4	0.281	0.319	0.343	0.360	0.403	0.369	0.543
lognormal	1.25	0.007	0.010	0.013	0.016	0.025	0.027	0.034
	1.5	0.028	0.038	0.046	0.053	0.073	0.068	0.097
	2	0.103	0.122	0.136	0.146	0.168	0.152	0.226
	4	0.351	0.363	0.369	0.373	0.373	0.348	0.509
pos. skewed	1.25	0.011	0.014	0.017	0.019	0.025	0.026	0.024
	1.5	0.048	0.055	0.060	0.063	0.064	0.061	0.065
	2	0.151	0.155	0.155	0.153	0.133	0.125	0.146
	4	0.396	0.379	0.363	0.349	0.285	0.264	0.335
neg. skewed	1.25	0.003	0.005	0.007	0.009	0.020	0.025	0.025
	1.5	0.009	0.013	0.018	0.023	0.052	0.061	0.069
	2	0.020	0.031	0.042	0.053	0.119	0.128	0.157
	4	0.061	0.093	0.125	0.157	0.318	0.277	0.361

7.2.1 Graf et al measure (1988) - $R^2_G(T^*)$

Table 7.1 displays that this measure is affected by the covariate distribution, it is reduced with negatively skewed covariates. This measure is an increasing function of the covariate effect, β, in all covariate distributions and time points. The measure tend to increase as the time point, T^*, increases in the normally and lognormally distributed covariates. Table 7.3 shows that the dispersion of the measure, as measured by the *C.V.*, decreases as the covariate effect, β, and T^* increase.

Table 7.2: Standard deviation of the sampling distribution of predictive accuracy measures and Schemper and Kaider's measure (1997) by the covariate distribution and covariate effect across all sample size conditions, censoring=0%

Covariate distribution	exp(β)	Graf et al measure at different time points					V_{SchH}	R^2_{SchK}
		$R^2_G(T_1)$	$R^2_G(T_2)$	$R^2_G(T_3)$	$R^2_G(T_4)$	$R^2_G(T_5)$		
normal	1.25	0.008	0.010	0.011	0.012	0.016	0.010	0.019
	1.5	0.016	0.019	0.020	0.022	0.025	0.017	0.030
	2	0.031	0.033	0.034	0.035	0.035	0.023	0.040
	4	0.053	0.048	0.045	0.044	0.041	0.026	0.040
lognormal	1.25	0.011	0.012	0.013	0.014	0.016	0.010	0.018
	1.5	0.023	0.024	0.025	0.025	0.025	0.016	0.029
	2	0.043	0.041	0.039	0.037	0.034	0.022	0.040
	4	0.056	0.050	0.047	0.045	0.041	0.025	0.041
pos. skewed	1.25	0.016	0.016	0.016	0.016	0.015	0.010	0.015
	1.5	0.034	0.031	0.029	0.027	0.021	0.015	0.024
	2	0.054	0.046	0.042	0.038	0.029	0.020	0.035
	4	0.061	0.055	0.050	0.048	0.038	0.025	0.044
neg. skewed	1.25	0.005	0.006	0.008	0.009	0.014	0.010	0.016
	1.5	0.008	0.010	0.012	0.013	0.021	0.016	0.026
	2	0.012	0.014	0.017	0.019	0.030	0.023	0.037
	4	0.020	0.024	0.028	0.031	0.043	0.027	0.046

7.2.2 Schemper and Henderson measure (2000) - V_{SchH}

Table 7.1 shows that this measure is an increasing function of the covariate effect, β. In the normally distributed covariate, the expected value of the measure varies from 0.027 to 0.369 for the range of β in the study. This measure is influenced by the covariate distribution, it decreases as the covariate distribution becomes asymmetric. Table 7.3 indicates that the dispersion of the measure decreases as the β increases in the model.

7.2.3 Schemper and Kaider measure (1997) - R^2_{SchK}

This measure is the only measure that does not belong to any of the proposed three main classes of predictive ability. In the normally distributed covariate, the expected value of the measure varies from 0.036 to 0.543 for the range of β in the study. The results of the simulation study in table 7.1 suggest that this measure is affected by the covariate distribution; it decreases as the covariate distribution becomes asymmetric. The measure increases as the covariate effect, β, becomes larger. Table 7.3 displays that the dispersion of the measure decreases with increasing covariate effect.

Table 7.3: Coefficient of variation of predictive accuracy measures and Schemper and Kaider's measure (1997) by the covariate distribution and covariate effect, expressed as %. Table entries are the average across all combinations of sample sizes, censoring=0%.

Covariate distribution	exp(β)	Graf et al measure at different time points					V_{SchH}	R^2_{ShK}
		$R^2_G(T_1)$	$R^2_G(T_2)$	$R^2_G(T_3)$	$R^2_G(T_4)$	$R^2_G(T_5)$		
normal	1.25	138.1	111.5	96.2	85.8	60.5	36.4	48.7
	1.5	79.4	63.7	54.5	47.8	33.1	23.1	27.5
	2	46.7	36.5	30.4	26.7	18.8	13.8	15.7
	4	17.7	14.3	12.6	11.5	9.7	6.6	7.0
lognormal	1.25	140.8	112.1	95.2	84.8	58.8	35.9	50.8
	1.5	77.6	60.5	51.1	44.4	31.9	22.7	28.8
	2	40.0	31.5	27.1	24.3	19.1	13.8	16.7
	4	15.2	13.1	12.1	11.4	10.4	6.9	7.7
pos. skewed	1.25	139.5	107.3	90.9	80.6	56.1	35.2	60.0
	1.5	67.4	53.6	46.1	41.3	31.8	23.0	35.9
	2	33.8	28.4	25.6	23.7	20.8	15.0	22.6
	4	14.7	13.7	13.1	12.9	12.7	8.9	12.6
neg. skewed	1.25	143.1	117.0	101.5	91.6	65.6	38.0	60.1
	1.5	86.4	69.7	60.4	54.0	38.8	25.6	35.9
	2	54.9	44.1	38.2	34.2	24.3	16.7	22.3
	4	30.7	24.9	21.4	18.8	12.8	9.4	12.1

7.3 Impact of censoring on the measures

In this section, we investigate the impact of censoring on $R^2_G(T^*)$, V_{SchH}, and R^2_{SchK} through a series of simulation studies similar to section 5.3. The results are summarised using similar methods to those of section 5.3.

Tables 7.4 to 7.6 summarise the results of the simulation studies. More detailed simulation results are presented in Appendix A. The tables in Appendix A summarise the impact of censoring by the covariate distribution, censoring type, and censoring proportion in a similar way to table 7.6. The figures in these tables are the average across four covariate effects, and three sample size conditions. It is evident from the tables that no summary statistic is presented for the Graf et al measure (1999) [31] in 80% censoring. We also presented the summary statistics for 4 time points, T_1 to T_4. To evaluate the predictive accuracy of a model using Graf et al measure (1999) [31], the time point of interest, T^*, should either be smaller than or equal to the last event time in the data. In some of the generated replicates, the time points, T_1 to T_5, were, by chance, larger than the last event time in small sample sizes, i.e. $n = 200$, when the amount of censoring was 80%. Although exponential distribution was used for the censoring distribution in random censoring condition, the risk set, i.e. time to the last event time, became shorter and shorter as the proportion of censored observations increased. Similar problem occurred

for $T_5 = 34.66$ in 50% censoring condition when the covariate effect was $\beta = 1.386$. This caused the program to stop, and we had to carry out the simulations with large sample size, i.e. 1,000, to investigate the performance of this measure in 80% censoring, which will be presented later in another table. We, therefore, present the summary statistic in 4 time points and 2 censoring proportions, i.e. 20% and 50% censoring, for the Graf et al's measure (1999).

Detailed simulation results are presented in Appendix A. The tables in Appendix A show the impact of censoring by the covariate distribution, censoring type, and censoring proportion in a similar way to table 7.4. The following sections describe the impact of censoring on each measure in details.

Table 7.4: The average percentage difference from the expected value of measures in the corresponding non-censored data by the covariate distribution and censoring proportion.

Covariate	%	Graf measure at different time points					
Distribution	Censored	$R^2_G(T_1)$	$R^2_G(T_2)$	$R^2_G(T_3)$	$R^2_G(T_4)$	V_{SchH}	R^2_{SchK}
normal	20	-1.3	-1.8	-2.2	-2.1	-0.6	-0.1
	50	-3.1	-4.5	-5.4	-5.4	-8.4	1.0
	80					-27.2	5.3
lognormal	20	-0.8	-1.3	-1.5	-1.4	1.0	-0.3
	50	-1.9	-3.2	-3.8	-3.4	-2.0	0.6
	80					-14.4	3.6
pos. skewed	20	-0.6	-0.8	-0.8	-0.5	2.8	-0.3
	50	-1.2	-1.9	-1.9	-1.1	7.9	0.2
	80					10.6	1.5
neg. skewed	20	-2.7	-3.4	-4.0	-4.1	-5.6	-0.4
	50	-5.9	-8.4	-10.1	-10.6	-21.8	1.7
	80					-44.0	11.4

7.3.1 Graf et al measure (1988) - $R^2_G(T^*)$

The simulation results indicate that this measure is almost unaffected by the amount of censoring in these experimental conditions. Table 7.4 shows that the average percentage change in the expected value of the measure is on average less than 10% in most of the experimental conditions. Table 7.6 show that the measure is unaffected by the amount of censoring in random censoring conditions, but decreases slightly in the type I or administrative censoring in all covariate distributions. The relative spread of the sampling distribution, indicated in table 7.5 , increases as the amount of censoring increases.

Due to computational issues in small samples as explained before, the performance of this measure in 80% censoring condition was not presented in the above tables. We,

Table 7.5: Coefficient of variation of measures by the covariate distribution and censoring proportion, expressed as %. Table entries are the average across three sample size conditions.

Covariate	%	Graf measure at different time points					
Distribution	Censored	$R^2_G(T_1)$	$R^2_G(T_2)$	$R^2_G(T_3)$	$R^2_G(T_4)$	V_{SchH}	R^2_{SchK}
normal	20	70.7	56.6	48.4	43.0	21.2	25.7
	50	72.3	57.7	49.2	43.8	23.8	29.9
	80					30.1	43.6
lognormal	20	68.8	54.4	46.5	41.3	21.1	26.6
	50	70.7	55.7	47.4	42.4	24.2	29.6
	80					31.2	39.0
pos. skewed	20	64.1	50.9	44.1	39.9	22.1	33.0
	50	65.6	52.0	45.2	41.4	28.0	33.9
	80					38.8	37.2
neg. skewed	20	78.8	63.8	54.9	49.1	24.2	33.5
	50	79.5	64.2	55.0	49.1	34.2	39.7
	80					48.4	65.1

therefore, carried out further simulations with large samples, $n = 1,000$, to examine this measure in the presence of heavy censoring. The simulations were run for two censoring proportions 0% and 80%, with 5,000 replicates in each experimental condition. Random non-informative right censoring was generated as described in section 4.3.4.

Table 7.7 shows the expected value and standard deviation of the sampling distribution of this measure evaluated at $T_2 = 8.13$. It is evident from the table that the expected value of this measure is consistent in the presence of heavy censoring across all covariate effects and covariate distributions.

7.3.2 Schemper & Henderson measure (2000) - V_{ShH}

Table 7.6 shows that this measure is not influenced by random censoring, except when the covariate distribution is negatively skewed, which decreases with the amount of censoring. However, the table suggests that there is an interaction between censoring and the covariate distribution in the type I or administrative censoring. The measure decreases on average in normal, lognormal, and negatively skewed distributions, but it increases in positively skewed distributions. Like Graf et al's measure (1999), the spread of the sampling distribution of the measure, expressed as the *C.V.* in table 7.5, increases with the amount of censoring.

Table 7.6: Summary performance of measures by the covariate distribution and censoring mechanism. Note that the entries for the Graf's measure (1999) do not include 80% censoring.

Measure		Covariate Distribution	Random Censoring Average % Difference	Random Censoring *C.V.*	Type I Censoring Average % Difference	Type I Censoring *C.V.*
Graf's measure (1999)	$R^2_G(T_1)$	normal	1.4	72.7	-5.7	70.3
		lognormal	1.7	70.9	-4.5	68.6
		pos. skewed	1.3	65.6	-3.1	64.1
		neg. skewed	-0.2	80.2	-8.4	78.0
	$R^2_G(T_2)$	normal	1.4	58.8	-7.7	55.4
		lognormal	1.4	56.4	-5.9	53.7
		pos. skewed	1.2	52.3	-3.8	50.6
		neg. skewed	0.3	66.3	-12.1	61.8
	$R^2_G(T_3)$	normal	1.1	50.9	-8.7	46.8
		lognormal	1.1	48.6	-6.4	45.3
		pos. skewed	1.1	45.6	-3.7	43.7
		neg. skewed	0.4	57.9	-14.5	51.9
	$R^2_G(T_4)$	normal	1.2	45.9	-8.7	40.9
		lognormal	1.1	43.6	-6.0	40.1
		pos. skewed	1.2	41.7	-2.7	39.6
		neg. skewed	1.2	52.8	-15.9	45.3
Schemper & Henderson (2000)	V_{SchH}	normal	-0.9	30.5	-23.3	19.6
		lognormal	1.5	28.6	-11.8	22.4
		pos. skewed	7.1	29.6	7.1	29.6
		neg. skewed	-23.9	35.5	-23.8	35.6
Schemper & Kaider (1997)	R^2_{ShK}	normal	2.0	33.3	2.1	32.8
		lognormal	1.4	32.1	1.3	31.3
		pos. skewed	0.5	34.9	0.4	34.5
		neg. skewed	3.9	45.8	4.6	46.4

7.3.3 Schemper & Kaider measure (1997) - R^2_{SchK}

This measure can be considered independent of censoring since the average percentage change in the expected value of the measure in both random and administrative censoring is less than 10% compared with the corresponding non-censored data (tables 7.4 and 7.6). The only exception is when the data is heavily censored, i.e. 80% censoring, and the covariate is heavily skewed to the left. In this case, the average percentage change in the expected value of the measure is 11.4% compared with the value of the measure in the corresponding non-censored data.

Table 7.7: The expected value and standard deviation (in brackets) of the sampling distribution of Graf et al (1999) measure in 0% and 80% censoring by the covariate effect and covariate distribution.

		Graf et al measure evaluated at $T_2 = 8.13$	
Covariate distribution	$\exp(\beta)$	0% censoring	80% censoring
normal	1.25	0.008 (0.006)	0.009 (0.007)
	1.5	0.027 (0.011)	0.028 (0.014)
	2	0.086 (0.020)	0.087 (0.025)
	4	0.319 (0.030)	0.319 (0.041)
lognormal	1.25	0.010 (0.007)	0.010 (0.009)
	1.5	0.037 (0.014)	0.038 (0.018)
	2	0.122 (0.025)	0.123 (0.031)
	4	0.363 (0.031)	0.364 (0.041
pos. skewed	1.25	0.013 (0.009)	0.015 (0.012)
	1.5	0.054 (0.019)	0.055 (0.022)
	2	0.155 (0.028)	0.155 (0.035)
neg. skewed	1.25	0.005 (0.004)	0.005 (0.005)
	1.5	0.013 (0.006)	0.014 (0.008)
	2	0.031 (0.009)	0.031 (0.011)

7.4 Consistency, distributional shape, and sample size effect

This section investigates the consistency and the shape of the sampling distribution of the measures as well as the impact of sample size.

7.4.1 Consistency of the measures

Both predictive accuracy measures are based on the measures of marginal and conditional prediction errors. In Graf et al's measure (1999) [31], $R^2_G(T^*)$, the prediction error is quantified by the average of the quadratic differences between an observed outcome, survival status, and the model-based survival probabilities, whereas in Schemper and Henderson measure (2000) [97] the prediction error is quantified by the average of the absolute differences of the same quantities.

In general, the marginal prediction error, D, is determined for a model without prognostic factors, and conditional prediction error, $D(X)$, is determined for a model with prognostic factors. Both measures provide a measure of predictive accuracy using $[D - D(X)]/D$ which evaluates the relative gain in predictive accuracy provided by the the prognostic factors when added to the model. A consistent estimator is the one whose estimators of marginal and conditional prediction errors are consistent.

In Graf's measure (1999) the marginal and conditional prediction errors at time T^*, $D(T^*)$ and $D_X(T^*)$, are defined as

$$D(T^*) = E\left[(Y(T^*) - S(T^*))^2\right]$$

and

$$D_X(T^*) = E\left[(Y(T^*) - S(T^*|X))^2\right]$$

where $Y(T^*)$ is the individual survival status at time T^*, i.e. equal to 0 if event happened before T^* and is equal to 1, otherwise. Graf (1998) [30] showed that $\widehat{D}_X(T^*)$ in equation 2.42 of chapter 2 and its marginal counterpart are consistent estimators of $D_X(T^*)$ and $D(T^*)$ in the Cox PH model.

In Schemper and Henderson's measure (2000), the corresponding population values of the marginal and conditional prediction errors, D_{SH} and $D_{SH}(X)$, are defined as

$$D_{SH} = 2\int_0^{T^*} S(t)\{1 - S(t)\}f(t)dt / \int_0^{T^*} f(t)dt$$

and

$$D_{SH}(X) = 2\int_0^{T^*} E_X\left[S(t|X)\{1 - S(t|X)\}\right]f(t)dt / \int_0^{T^*} f(t)dt$$

where $[0, T^*)$ is the follow-up period. Schemper and Henderson (2000) [97] showed that the estimator of conditional prediction error, $\widehat{D}_{SH}(X)$, equation 2.43 of chapter 2, and its marginal counterpart, $\widehat{D}_{SH}$, are a consistent estimator of the $D_{SH}(X)$ and D_{SH}.

Finally, Schemper and Kaider's measure (1997) provides a non-parametric measure of correlation.

7.4.2 Sampling distribution of the measures

Figure 7-1 depicts the sampling distribution of the Schemper and Henderson (2000) [97], V_{SchH}, and Schemper and Kaider (1997) [98], R^2_{SchK}, measures from the simulation studies. In the simulations, the covariate is normally distributed with $5,000$ replicates in each experimental condition. The survival times are randomly censored by considering an exponential distribution for censoring times, as described in section 4.3.4. The shape of the sampling distribution of both measures are similar to those of explained variation and randomness measures in chapters 5 and 6.

The sampling distribution of the Schemper and Henderson (2000) [97] confirms the consistency of the estimator, $\widehat{V}_{SchH}$ as defined in equation 2.44, in random censoring condition. All distributions in figure 7-1 tend towards a spike over the parameter of interest as n becomes ever larger, as those of Schemper and Kaider measure [98], R^2_{SchK}. The sampling distribution of both measures exhibit considerable skewness, particularly when the covariate effect is small and censoring is more than 50%. The shape of the sampling distribution of Graf et al measure (1999) is similar to those of Schemper and Henderson (2000) [97], V_{SchH}.

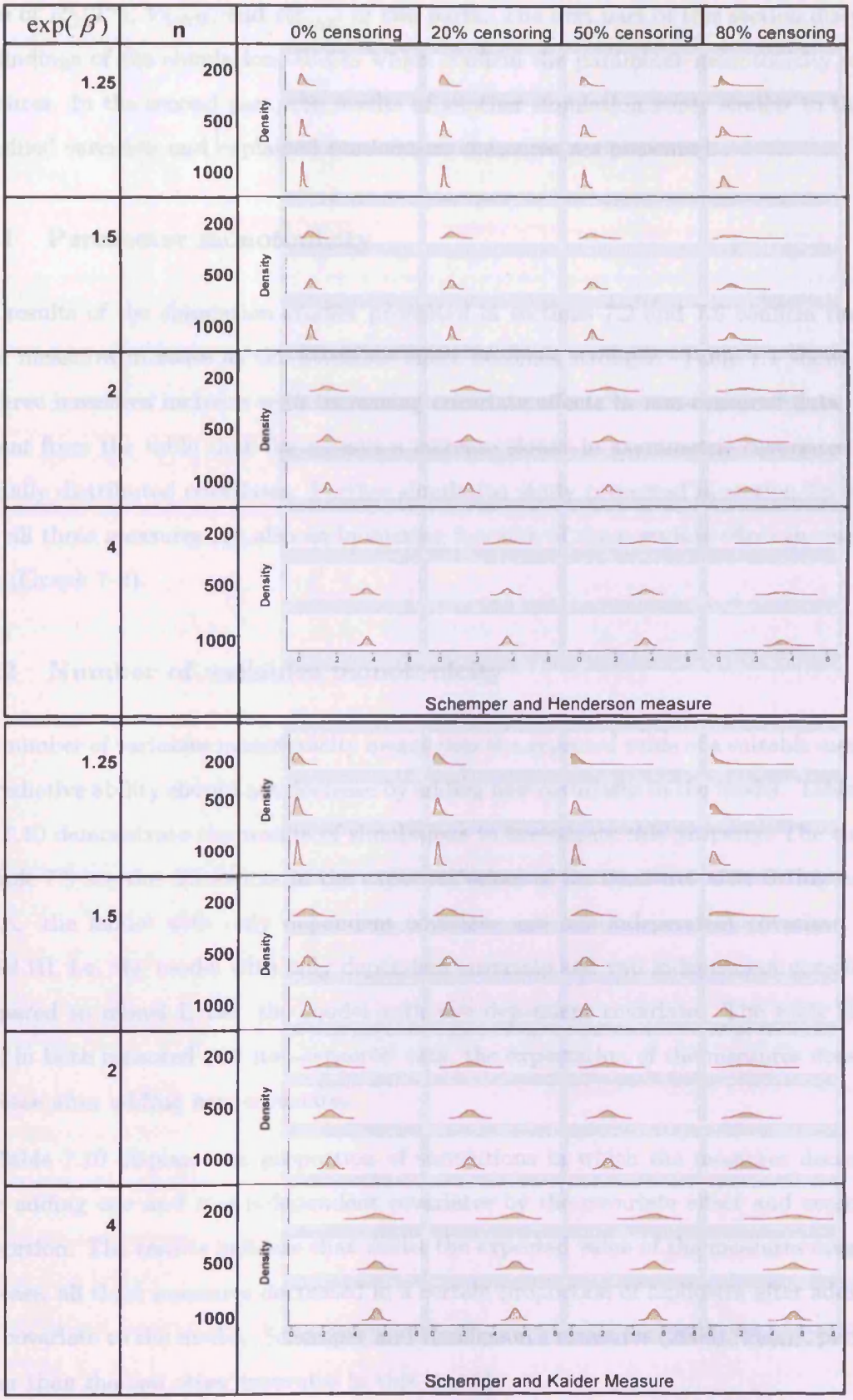
7.4.3 Impact of sample size on the measures

In a similar simulation study to those of explained variation and explained randomness measures, we evaluated the impact of sample size on the three measures studied in this chapter. The results are tabulated in table 7.8 which indicate that the measures increase slightly when the effective sample size, i.e. number of events, is small. The table shows that the measures increase when both sample size and the covariate effect are small, i.e. $n = 200$ and $\exp(\beta) = 1.25$, and the amount of censoring is high, i.e. 80%. We observed a similar pattern in other simulation studies when we studied skewed covariates and a different censoring mechanism, i.e. type I or administrative censoring.

Table 7.8: Percentage change in the expected value of measures in small and large sample sizes by censoring proportion. The figures in brackets are the standard deviation of the sampling distribution.

		20% Censoring			80% Censoring		
		Sample size			Sample size		
Measure	$\exp(\beta)$	1000	200	% Change	1000	200	% Change
$R^2_G(T_1)$	1.25	0.005	0.006	20%	0.006	0.008	33%
		(0.005)	(0.012)		(0.006)	(0.016)	
	4	0.282	0.282	0%	0.283	0.283	0%
		(0.032)	(0.072)		(0.039)	(0.089)	
V_{SchH}	1.25	0.026	0.028	7%	0.026	0.032	22%
		(0.007)	(0.016)		(0.015)	(0.032)	
	4	0.37	0.368	0%	0.347	0.337	-3%
		(0.017)	(0.037)		(0.031)	(0.064)	
R^2_{SchK}	1.25	0.035	0.039	12%	0.037	0.049	31%
		(0.035)	(0.039)		(0.037)	(0.049)	
	4	0.543	0.539	-1%	0.548	0.545	-1%
		(0.543)	(0.539)		(0.548)	(0.545)	

Figure 7-1: Sampling distributions of Schemper and Henderson (2000) and Schemper and Kaider (1997) measures by the covariate effect, sample size, and censoring proportions in the normally distributed covariate and random censoring conditions.



7.5 Monotonicity property of proposed measures

In this section, we investigate the parameter and number of variables monotonicity properties of $R^2_G(T^*)$, V_{SchH}, and R^2_{SchK} in two parts. The first part of this section discusses the findings of the simulation studies which confirm the parameter monotonicity of the measures. In the second part, the results of another simulation study similar to that of explained variation and explained randomness measures are presented.

7.5.1 Parameter monotonicity

The results of the simulation studies presented in sections 7.2 and 7.6 confirm that all three measures increase as the covariate effect becomes stronger. Table 7.1 shows that all three measures increase with increasing covariate effects in non-censored data. It is evident from the table that the measures increase slower in asymmetric covariates than normally distributed covariates. Further simulation study presented in section 7.6 shows that all three measures are also an increasing function of the covariate effect in censored data (Graph 7-2).

7.5.2 Number of variables monotonicity

The number of variables monotonicity means that the expected value of a suitable measure of predictive ability should not decrease by adding new covariates to the model. Tables 7.9 and 7.10 demonstrate the results of simulations to investigate this property. The entries in table 7.9 are the differences in the expected values of the measures after fitting model II, i.e. the model with only dependent covariate and one independent covariate, and model III, i.e. the model with only dependent covariate and two independent covariates, compared to model I, i.e. the model with one dependent covariate. The table shows that, in both censored and non-censored data, the expectation of the measures does not decrease after adding new covariates.

Table 7.10 displays the proportion of simulations in which the measures decreased after adding one and two independent covariates by the covariate effect and censoring proportion. The results indicate that whilst the expected value of the measures does not decrease, all three measures decreased in a certain proportion of replicates after adding a new covariate to the model. Schemper and Henderson's measures (2000), V_{SchH}, perform better than the two other measures in this regard.

Table 7.9: Mean difference in the expected value of measures after adding one or two independent covariates to the model in 2,000 simulations, normally distributed covariates.

Measure	$\exp(\beta)$	Model II		Model III	
		Mean difference to model I		Mean difference to model I	
		non-censored	censored	non-censoring	censored
$R^2_G(T_2)$	1.25	0.000	0.001	0.001	0.001
	1.5	0.000	0.001	0.001	0.001
	2	0.000	0.001	0.001	0.002
	4	0.001	0.001	0.001	0.002
V_{SchH}	1.25	0.001	0.004	0.002	0.008
	1.5	0.001	0.004	0.002	0.007
	2	0.001	0.003	0.002	0.006
	4	0.001	0.002	0.001	0.004
R^2_{SchK}	1.25	0.001	0.007	0.003	0.013
	1.5	0.001	0.006	0.003	0.012
	2	0.001	0.005	0.002	0.009
	4	0.001	0.002	0.001	0.005

7.6 Upper bound of the measures

In this section, we demonstrate the upper bound of the measures by applying similar simulation studies to section 5.6. In the simulations, survival times are exponentially distributed, the covariate is normally distributed $X \sim N(0,1)$, sample size is 500, and non-informative random censoring was generated by considering an exponential distribution for the censoring times with $2,000$ replicates in each experimental condition. For the Graf et al's measure (1999), we have carried out the simulations in three time points $T_1 = 5.27$, $T_2 = 8.13$, and $T_3 = 11.16$.

A comparison of simulation results in section 7.2 with the corresponding sections of chapters 5 and 6 clarifies that the predictive accuracy measures, $R^2_G(T^*)$ and V_{SchH}, attain lower values than the explained variation and explained randomness measures. To examine whether these measures reach values close to 1 in theory, we carried out the simulations for a wider range of covariate effect from, i.e. $\beta = 0.22$ ($\exp(\beta) = 1.25$) to $\beta = 8.32$ ($\exp(\beta) = 4096$), than we did for the explained variation and explained randomness measures. However, hazards ratios of this magnitude, i.e. $HR = 4096$, are rare in practical applications.

Figure 7-2 displays the expected value of measures from $\beta = 0.22$ ($\exp(\beta) = 1.25$) to $\beta = 8.32$ ($\exp(\beta) = 4096$) in 0% and 50% censoring conditions. All three measures are an increasing function of the covariate effect. With increasing covariate effect, the predictive accuracy measures, $R^2_G(T^*)$ and V_{SchH}, increase slower than Schemper and

Table 7.10: Proportion decrease in measures after adding independent covariate(s) to the model in 2000 simulations, normally distributed covariates.

Measure	exp(β)	Model II		Model III	
		Prop. decreased to model I		Prop. decreased to model I	
		non-censored	censored	non-censoring	censored
$R^2_G(T_2)$	1.25	0.44	0.40	0.41	0.35
	1.5	0.43	0.41	0.41	0.37
	2	0.42	0.38	0.37	0.35
	4	0.41	0.38	0.38	0.34
V_{SchH}	1.25	0.17	0.20	0.07	0.09
	1.5	0.16	0.19	0.07	0.09
	2	0.15	0.22	0.06	0.12
	4	0.14	0.25	0.05	0.15
R^2_{SchK}	1.25	0.32	0.15	0.23	0.06
	1.5	0.32	0.19	0.23	0.08
	2	0.34	0.21	0.25	0.12
	4	0.36	0.30	0.27	0.20

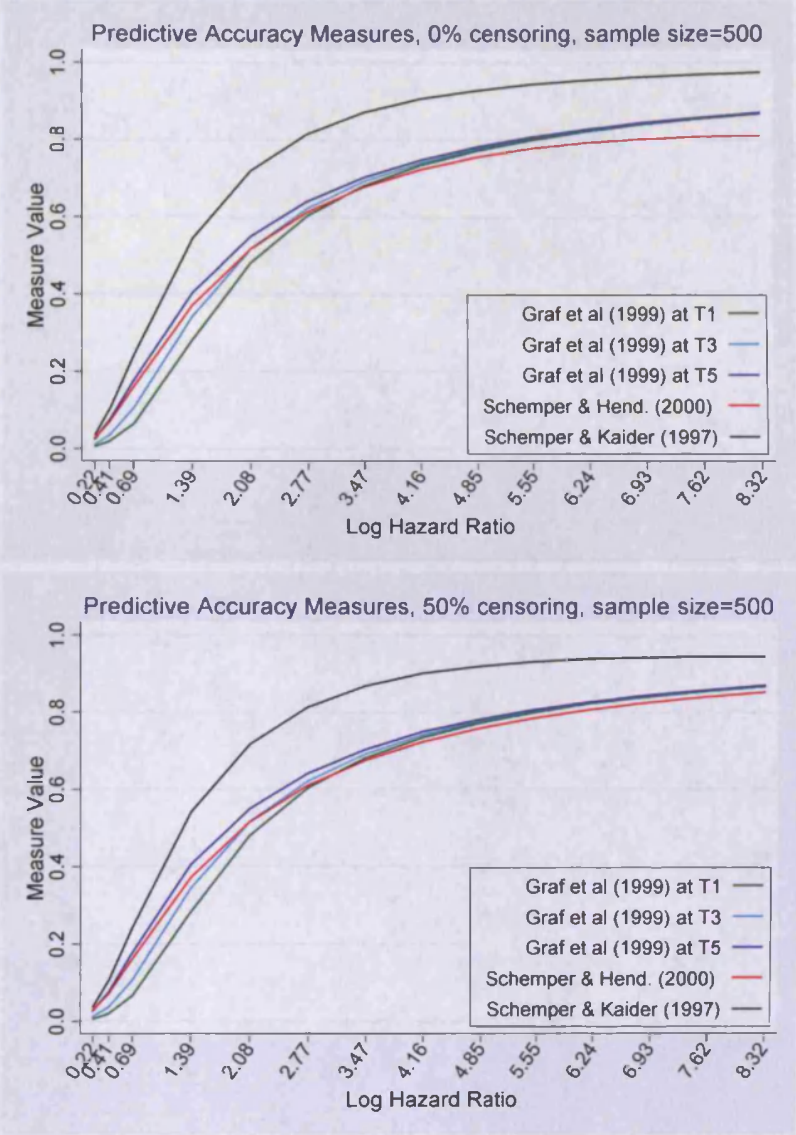
Kaider measure (1997) [98], R^2_{SchK}.

Both predictive accuracy measures, $R^2_G(T^*)$ and V_{SchH}, are in agreement for the range of the covariate effect studied here. It is evident from figure 7-2 that the expected value of Graf et al's measure (1999) in the three time points converges as the covariate effect becomes larger; they all reach values near 0.90. The expected value of Schemper and Henderson measure (2000), V_{SchH}, increases in both censored and non-censored data. It appears that this measure levels off after $\beta = 6.24$ ($HR = 512$) in non-censored data, whereas it still increases in censored data. Schemper and Kaider's measure (1997) [98] increases rapidly with increasing covariate effect and reaches values near 1 for large but reasonable covariate effects.

7.7 Robustness of the measures

Simulations were carried out to investigate the impact of extreme and outlier observations on the three measures investigated in this chapter. This section consists of two parts which demonstrate the impact of extreme and outlier observations on the measures of predictive accuracy, respectively. We show the results of an simulation study carried out for the covariate effect $\beta = 0.69$, sample size $= 200$, and 50% censoring condition with $2,000$ replicates in each experimental condition. We contaminated the data sets with extreme and outlier observations in the same way as we did for the study on the robustness of the explained variation measures, which was described in section 5.7, and present the results

Figure 7-2: Measures as a function of covariate effect in the model, normally distributed covariate. In the bottom graph, survival times are randomly censored according to an exponential distribution for censoring times.



through similar graphs.

7.7.1 Impact of extreme observations

Graph 7-3 displays the expected value of measures as one observation in the data set becomes more extreme. If a measure is resistant to extreme observations, the curve which represents the measure is expected to be a flat line across the X axis. The graph demonstrates that the measures are resistant to extreme observations since the expected value of the measures remain relatively constant as one of the observations becomes more extreme, i.e. the covariate and corresponding outcome value, i.e. time, move towards the

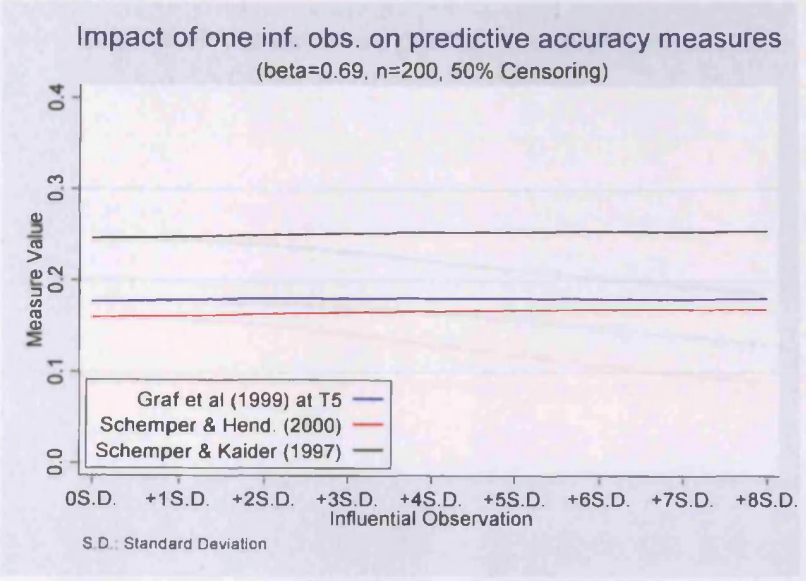


Figure 7-3: Mean of the sampling distribution of two predictive accuracy measures and Schemper and Kaider measure (1997) as the extreme observation becomes more influential.

extremes of their respective distributions.

7.7.2 Impact of outlier observations

Graph 7-4 displays the results of a similar simulation study to show the impact of outlier observations. Similar to graph 7-3, we expect flat lines across the X axis if the measures are resistant to such observations. The graph demonstrates that the measures are influenced by the outliers in the data set.

Limited simulation studies were carried out for other experimental conditions which showed that, in general, outlier observations have more impact on the measures in small sample sizes than the large ones.

7.8 Impact of model mis-specification on the measures

This section investigates the effect of model mis-specification on $R^2_G(T^*)$, V_{SchH}, and R^2_{SchK}. This section consists of three parts, each examining the impact of under-fitting, covariate mis-modelling, and non-proportional hazards on the measures. Simulation studies similar to those of section 5.8 were carried out to study the issue of model mis-specification on the measures; therefore, we do not describe the study design in this

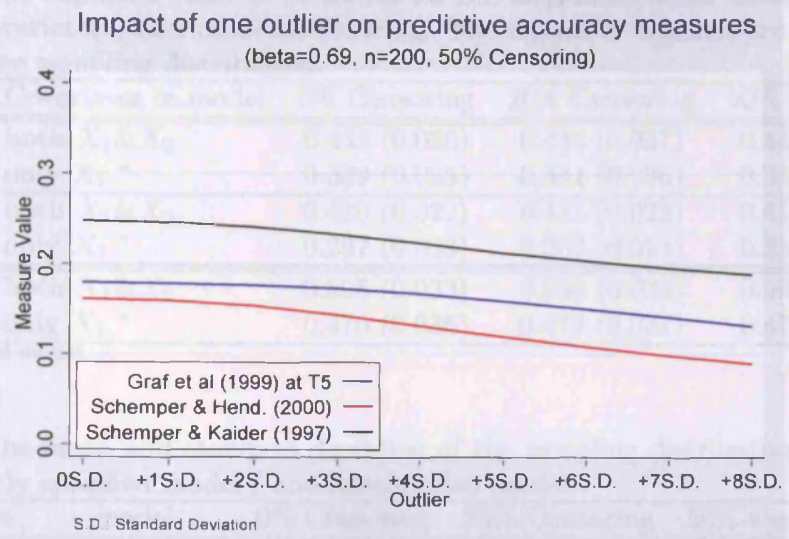


Figure 7-4: Mean of the sampling distribution of two predictive accuracy measures and Schemper and Kaider measure (1997) as the outlier observation becomes more influential.

section again. All the simulations were carried out in different censoring conditions with 500 sample size and 2,000 replicates in each experimental condition. The results are summarised in similar tables to those of section 5.8.

7.8.1 Impact of under-fitting - covariate omission

Table 7.11 demonstrates the impact of under-fitting on $R^2_G(T^*)$, V_{SchH}, and R^2_{SchK}. The table presents the expected value and standard deviation of the sampling distribution of the measures for the full and under-fitted models by the amount of censoring. Unlike the explained variation and explained randomness measures, where under-fitting imposes further bias on the measures in censored data, the expected value of the measures studied in this section remain relatively constant in the under-fitted model across different censoring proportions. The dispersion of the measures in both full and under-fitted models increase as the amount of censoring increases.

7.8.2 Impact of covariate mis-modelling

In a similar simulation study to those of explained variation and explained randomness measures, we examine the impact of covariate mis-modelling on $R^2_G(T^*)$, V_{SchH}, and R^2_{SchK} in this section. We repeated the studies to investigate the impact of modelling the

Table 7.11: The expected value of measures for full and under-fitted models. Normally distributed covariate(s) and random censoring. The figures in brackets are the standard deviation of the sampling distribution.

Measure	Covariates in model	0% Censoring	20% Censoring	50% Censoring
$R^2_G(T_5)$	both X_1&X_2	0.443 (0.036)	0.444 (0.037)	0.445 (0.046)
	only X_1 *	0.339 (0.035)	0.341 (0.036)	0.342 (0.046)
V_{SchH}	both X_1&X_2	0.410 (0.022)	0.411 (0.023)	0.410 (0.026)
	only X_1 *	0.297 (0.023)	0.302 (0.024)	0.311 (0.028)
R^2_{SchK}	both X_1&X_2	0.595 (0.033)	0.596 (0.034)	0.593 (0.036)
	only X_1 *	0.470 (0.036)	0.473 (0.037)	0.478 (0.040)

*=under-fitted model

Table 7.12: The mean and standard deviation of the sampling distribution of measures for the correctly specified model I and misspecified model.

Measure	model	0% Censoring	20% Censoring	50% Censoring
$R^2_G(T_5)$	true model I	0.268(0.033)	0.268(0.035)	0.269(0.042)
	missp. model	0.264(0.034)	0.263(0.036)	0.263(0.045)
V_{SchH}	true model I	0.247(0.022)	0.247(0.023)	0.251(0.029)
	missp. model	0.239(0.023)	0.245(0.025)	0.263(0.031)
R^2_{SchK}	true model I	0.365(0.037)	0.365(0.039)	0.363(0.046)
	missp. model	0.365(0.037)	0.370(0.039)	0.388(0.044)

covariate, X, as linear function of log hazard ratio in the Cox PH model where the true functional form of the covariate is either "model I", i.e. $f_1(X)$, where

$$f_1(X) = 0.932 * X + 0.156 * X^2 + 0.014 * X^3$$

or "model II", i.e. $f_2(X)$, where

$$f_2(X) = 0.668 * X - 0.413 * X^2 + 0.045 * X^3$$

where X is normally distributed, $X \sim N(0, 1)$. Figure 5-7 of chapter 5 illustrates the relationship between both true and linear models and the log hazards ratio.

Model I

Table 7.12 shows the expected value and standard deviation of the sampling distribution of the measures for true and mis-specified models by censoring proportions. The table indicates that the measure proposed by Graf et al (1999), $R^2_G(T^*)$ is consistent in both true and mis-specified models, whereas V_{SchH} and R^2_{SchK} increase in the mis-specified model as the amount of censoring increases.

Table 7.13: The mean and standard deviation of the sampling distribution of measures for the correctly specified model II and misspecified model.

Measure	model	0% Censoring	20% Censoring	50% Censoring
$R^2_G(T_5)$	true model II	0.134(0.026)	0.134(0.027)	0.136(0.031)
	missp. model	0.086(0.026)	0.088(0.026)	0.090(0.029)
V_{SchH}	true model II	0.193(0.023)	0.192(0.023)	0.188(0.027)
	missp. model	0.128(0.022)	0.123(0.022)	0.111(0.023)
R^2_{SchK}	true model II	0.237(0.037)	0.237(0.038)	0.237(0.041)
	missp. model	0.220(0.037)	0.213(0.037)	0.191(0.037)

Model II

Similarly, table 7.13 shows the mean and standard deviation of the sampling distribution of the measures for the true and mis-specified models by censoring proportions. Similar conclusions can be drawn for this case, except that the measures V_{SchH} and R^2_{SchK} decrease in the mis-specified model as the amount of censoring increases.

7.8.3 Non-proportional hazards

In an analogous simulation study to those of explained variation measures in section 5.8.3, we examined the impact of non-proportional hazards on $R^2_G(T^*)$, V_{SchH}, and R^2_{SchK}. Simulation results under non-proportional hazards are displayed in table 7.14. The entries of the table represented in *italics* are the expected value of the measures when the hazard ratio does not change (i.e. $HR1 = HR2 = 0.5$), i.e. when the proportional hazards assumption holds. In this case, $R^2_G(T_5)$ and V_{SchH} agree and R^2_{SchK} results in slightly higher values. Furthermore, the impact of non-proportional hazards on the measures diminishes as the amount of censoring increases.

Table 7.14: Simulation results for non-proportional hazards. HR1 - hazard ratio in favour of treatment arm for the first two years in trial. HR2 - hazard ratio after two years in trial. Sample size is 500 in all experimental conditions, and survival times are randomly censored. The standard deviations are given in brackets

Measure	$HR1$	$HR2$	0% Cens.	20% Cens.	50% Cens.	80% Cens.
$R^2_G(T_5)$	0.5	0.1	0.051(0.027)	0.056(0.029)	0.068(0.034)	
	0.5	0.3	0.065(0.025)	0.067(0.027)	0.070(0.034)	
	0.5	*0.5*	*0.068*(0.023)	*0.069*(0.025)	*0.070*(0.033)	
	0.5	0.7	0.066(0.021)	0.067(0.024)	0.070(0.033)	
	0.5	0.9	0.062(0.020)	0.065(0.023)	0.070(0.032)	
V_{SchH}	0.5	0.1	0.091(0.017)	0.089(0.018)	0.072(0.020)	0.051(0.027)
	0.5	0.3	0.069(0.015)	0.068(0.016)	0.064(0.020)	0.052(0.028)
	0.5	*0.5*	*0.055(0.013)*	*0.056(0.014)*	*0.056(0.019)*	*0.052(0.029)*
	0.5	0.7	0.046(0.012)	0.047(0.013)	0.050(0.018)	0.053(0.030)
	0.5	0.9	0.038(0.011)	0.040(0.012)	0.046(0.017)	0.053(0.031)
R^2_{SchK}	0.5	0.1	0.118(0.028)	0.116(0.029)	0.103(0.032)	0.089(0.047)
	0.5	0.3	0.098(0.025)	0.095(0.025)	0.093(0.031)	0.089(0.046)
	0.5	*0.5*	*0.085(0.024)*	*0.086(0.024)*	*0.086(0.030)*	*0.089(0.046)*
	0.5	0.7	0.076(0.023)	0.080(0.024)	0.080(0.030)	0.089(0.046)
	0.5	0.9	0.069(0.022)	0.070(0.024)	0.076(0.029)	0.088(0.046)

7.9 Discussion

In this chapter, we studied the predictive accuracy measures, $R^2_G(T^*)$ and V_{SchH}, proposed by Graf et al (1999) [31] and Schemper and Henderson (2000) [97], and the measure proposed by Schemper and Kaider (1997) [98], R^2_{SchK}. We carried out similar simulation studies to those of explained variation and explained randomness measures, presented in chapters 5 and 6, to compare all the measures consistently.

The results of our simulation studies in section 7.2 imply that $R^2_G(T^*)$ and V_{SchH} are generally lower than the explained variation and explained randomness measures. The results of our simulation studies in section 7.2 showed that the expected value of R^2_G depends on the time point that is used to evaluate predictive accuracy. If the time point of interest is at the beginning of the study where survival probabilities are near 1, we observe less variability and eventually low predictive ability. The measure proposed by Schemper and Kaider (1997) [98], R^2_{SchK}, is in agreement with the explained variation measures if the distribution of the covariate is either symmetric or moderately asymmetric, i.e. normal and lognormal distributions.

The results of our simulation studies on the impact of censoring in section 7.3 indicate that $R^2_G(T^*)$ and R^2_{SchK} are largely unaffected by the amount of censoring. The measure V_{SchH} performs well in random censoring with symmetric or moderately asymmetric covariate distributions, otherwise it is affected by the amount of censoring. Table 7.15 summarises the findings of our simulation studies carried out to investigate the impact of censoring on the measures. The codes in the table show the extent of censoring effect on the measures, with 1 representing almost no effect, i.e. the average percentage change in the mean of sampling distribution is $0\% - 9\%$ compared with the expected value of the measure in the corresponding non-censored data, and 4 representing a large effect, i.e. the average percentage change in the mean of sampling distribution is over 50% (compared with the expected value of the measure in the corresponding non-censored data.) The tables indicate that $R^2_G(T^*)$ performs reasonably well with respect to censoring in all experimental conditions.

Consistency and the sampling distribution of measures were investigated in section 7.4. Our investigation found that the measures are consistent in the presence of random censoring. The sampling distribution of the Schemper and Henderson (2000) and Schemper and Kaider (1997) measures were presented in figure 7-1 for different covariate effects and censoring proportions. Similar to the measures investigated in chapters

Table 7.15: Summary of censoring effects on predictive accuracy and Schemper and Kaider (1997) measures by the covariate distribution and censoring type. The codes show the extent of censoring effect in different situations from almost no effect, 1, to a large effect, 4.

Censoring type	P. A. measure	Covariate or Prognostic Index Distribution			
		Normal	Lognormal	Pos. skewed	Neg. skewed
Random censoring	$R^2_G(T^*)$	1	1	1	1
	V_{SchH}	1	1	1	3
	R^2_{SchK}	1	1	1	1
Type I censoring	$R^2_G(T^*)$	1	1	1	2
	V_{SchH}	3	2	1	3
	R^2_{SchK}	1	1	1	1

1: Almost no effect, i.e. the average percentage change in the mean of sampling distribution is 0%−9%

2: Slight effect, i.e. the average percentage change in the mean of sampling distribution is 10%−19%

3: Moderate effect, i.e. the average percentage change in the mean of sampling distribution is 20%−49%

4: Large effect, i.e. the average percentage change in the mean of sampling distribution is over 50%

5 and 6, the sampling distributions of both estimators show considerable skewness when censoring is more than 50%.

Sample size has a moderate effect on the measures if both number of events and covariate effect are small. Table 7.8 indicate that the measures increase by about 22% − 33% in these circumstances.

Graphs presented in section 7.6 and tables 7.9 and 7.10 indicate that the measures satisfy both monotonicity properties. Furthermore, the investigation which was carried out in section 7.7 to examine the impact of extreme and outlier observations shows that the measures perform satisfactorily in the presence of extreme values, but they decrease in the presence of outlier observations. For example, in the presence of severe outlier observations, i.e. $m = 8$ in section 7.7, $R^2_G(T^*)$, V_{SchH}, and R^2_{SchK} decrease by about 34%, 44%, and 23% respectively (the expected value of the measures at $8SD$ are compared with the corresponding values at $0SD$, no contamination, in figure 7-4), whereas they increase by only 6%, 4%, and 2% in the presence of influential extreme observations (figure 7-3).

The predictive accuracy measures $R^2_G(T^*)$ and V_{SchH}, can reach high values, i.e. more than 0.80, in theory. The graphs in section 7.6 suggested that we need strong prognostic factors to be able to predict the individual's status as "dead" or "alive" using the survival models. The measure proposed by Schemper and Kaider (1997), R^2_{SchK}, also reaches values near 1.

Finally, the impact of three types of model mis-specification was investigated in section

Table 7.16: Summary of sample size effect and parameter monotonicity of predictive accuracy and Schemper and Kaider (1997) measures.

Measure	Sample Size	Does parameter monotonicity hold?
$R^2_G(T^*)$	no effect[1]	yes[2]
V_{SchH}	no effect[1]	yes[2]
R^2_{SchK}	no effect[1]	yes[2]

1) There is a moderate effect of sample size on measures only when covariate effect is 1.25, sample size is 200, and censoring proportion is high, i.e. 80%.

2) The measure increases with increasing parameter effect.

7.8. Unlike the explained variation and explained randomness measures, $R^2_G(T^*)$ results in consistent values under different degrees of censoring in models that are either under-fitted or their covariate is mis-modelled. In under-fitted models, the measures V_{SchH} and R^2_{SchK} are consistent under different degrees of censoring. They, however, are inconsistent if the covariate is mis-modelled. Similar to the impact of non-proportional hazards on the explained variation and explained randomness measures, the susceptibility of the measures to non-proportional hazards diminishes as the amount of censoring increases.

In summary, the measure proposed by Graf et al (1999), $R^2_G(T^*)$, performs reasonably well with regard to the essential properties outlined in chapter 3. It is unaffected by the amount of censoring, is consistent, and satisfies the monotonicity properties. Moreover, it results in consistent values in the case of model mis-specification. However, this measure evaluates the predictive ability of the model at a specific time point, and its value changes with the time point of interest. The alternative measure, V_{SchH}, provides an overall measure of predictive accuracy. This measure performs well in the case of random censoring when the covariate is symmetric or moderately asymmetric. Between the two measures, $R^2_G(T^*)$ is preferred if we can not rely on the model. The measure proposed by Schemper and Kaider (1997), R^2_{SchK}, performs well with regard to the essential properties.

In the last three chapters, we carried out simulation studies to investigate the proposed measures of predictive ability in survival models. In the next chapter, we apply them to the data sets from real studies.

Chapter 8

Applications to medical research and data analysis

8.1 Introduction

In this chapter, we apply the potentially recommendable measures of predictive ability discussed in the last three chapters to real data. We compute the measures for the proposed survival models for different diseases. The aims of this chapter are:

I) to illustrate the applications of the predictive ability measures in medical research

II) to quantify the predictive ability of available/known prognostic factors

III) to compare the measures in each category systematically in real data sets

IV) to explain the observed discrepancies in the estimates of proposed measures based on the results of our investigations in the previous chapters.

This chapter consists of 3 sections. First, a summary of the data sets and the proposed regression models are presented. The data sets are chosen from different diseases to examine the performance of the proposed measures in various disease types. The data sets have a wide range of censoring proportions and sample size conditions.

In the second section, we present the estimates of predictive ability measures with the corresponding bootstrap confidence intervals. We apply survival models from literature to these data sets. Some of the proposed models have been developed to study the impact of a particular treatment on the survival of patients, while others have been developed as

a prognostic model. An important characteristic of prognostic models is their consistency with basic medical knowledge.

Multivariable fractional polynomial (MFP) approach, introduced by Royston and Altman (1994) [89], is a method which ensures that the resulting models are both parsimonious and consistent with basic medical knowledge. It is a strategy in which continuous predictors are kept continuous, and nonlinear relationships (if present) are detected and modelled appropriately. We apply this approach to the data sets and compare the predictive ability of the models based on fractional polynomial approach to other proposed models.

In medical research, continuous variables are often converted into categorical variables by grouping values into two or more categories. In some proposed models, continuous prognostic factors, such as age, are introduced into multivariable regression models as categorical variables. We identify these models as "linear models" in our studies and compare their predictive ability to models developed using fractional polynomial approach. We only report the estimated predictive ability measures in this chapter; the estimated coefficients and goodness of fit measures of the proposed models are presented in Appendix C. We discuss the findings based on the results of the simulation studies in the previous chapters. Finally, a discussion of the main points is presented.

8.2 Clinical data sets

In this section, we give a summary of 9 data sets that we use to describe the predictive ability measures using real data. The data sets are mainly from clinical trials in breast, renal, and prostate cancers, and diseases such as leg ulcer and primary biliary cirrhosis (PBC). The data sets are from studies that were generally carried out by research organisations to investigate the impact of the prognostic factors on the survival of patients in the relevant disease types.

8.2.1 Data set 1: venous leg ulcer

The first data set is from a clinical trial which was carried out to evaluate prognostic factors in uncomplicated venous leg ulcer healing (Smith et al, 1992 [106]). The data consists of several covariates and one outcome variable on 200 individuals. The covariates are clinical and biological factors such as age, diastolic blood pressure, height, ankle

pressure, body weight, presence or absence of deep vein involvement, and treatment differences. The outcome variable, survival time, is the number of days from diagnosis to complete healing.

Smith et al (1992) [106] fitted a Cox proportional hazards regression model to investigate the prognostic factors in this study. Royston and Altman (1994) [89] discussed two models, MFP I & MFP II in tables 8.1 to 8.8, based on a multivariable fractional polynomial (MFP) approach [89]. The MFP algorithm resulted in a model, MFP I, which contain five prognostic factors as significant at the 5% level in a multivariable model as age, months since onset, initial ulcer area (mm^2), diastolic blood pressure (mm Hg), and deep vein involvement ($1 = Y, 0 = N$). In this model, the covariates age, months since onset, and initial ulcer area (mm^2) were subject to an FP1 transformation with powers -2, 0, and 0.5. Royston and Altman (1994) [89] suggested an alternative model, MFP II, which is biologically more plausible. In this model, only months since onset was subject to FP1 transformation with power 0. The estimated coefficients and goodness of fit measures of both models are presented in Appendix C.

8.2.2 Data set 2: breast cancer I

The second data set is a sample of 295 women with breast cancer (Van't Veer et al (2002) [112]). Van't Veer et al (2002) [112] used this data set to develop a 70-gene classifier to predict survival in young patients with stage I or stage II breast cancer. The gene-expression data set was derived by researchers from the Netherlands Cancer Institute and Rosetta Inpharmatics–Merck using oligonucleotide microarrays (Agilent). Data on recurrence-free survival (RFS), defined as the time to a first event, and overall survival (OS) were available for all patients. Most of the patients had stage I or II breast cancer; 165 had received local therapy alone, 20 had received tamoxifen only, 20 had received tamoxifen plus chemotherapy, and 90 had received chemotherapy only.

Cheng Fan et al (2006) [25] analysed this data set further and fitted different multivariable Cox proportional hazards models using recurrence-free survival (RFS) and overall survival (OS) as two different end points. They first included clinical prognostic factors alone in the models, then "70-gene predictor" was added to the model to evaluate the its effect on RFS and OS. We identify the models containing only the biological prognostic factors as RFS I and OS I, and the models containing both the biological prognostic factors and the 70-gene predictor as RFS II and OS II in tables 8.1 to 8.8. The mod-

els including only biological prognostic factors comprise age (as a continuous variable), oestrogen-receptor status (positive vs. negative), tumour grade (1 vs. 2 and 1 vs. 3), nodal status (no positive nodes vs. one to three positive nodes and no positive nodes vs. more than three positive nodes), tumour diameter (2 cm or less vs. more than 2 cm), and treatment received (no adjuvant therapy vs. chemotherapy, hormonal therapy, or both). Models RFS II and OS II contain 70-gene predictor as well. We display the estimated coefficient and goodness-of-fit measures, which were included in the supplementary material in Cheng Fan's (2006) paper [25], in Appendix C.

8.2.3 Data set 3: breast cancer II

The third data set is from German Breast Cancer Study Group which carried out a comprehensive cohort study in primary nodes positive breast cancer [102]. Randomised and non-randomised patients were eligible, and about two-thirds were entered into the randomised part. This study recruited 720 individuals of which 686 had complete information, of which 299 experienced the event of interest (RFS). Besides treatment, data on other clinical and biological factors such as age, tumour size, number of lymph nodes, progesterone and oestrogen respecter status, menopausal status, and tumour grade were collected.

The aim of the study was to investigate the prognostic factors in node positive breast cancer and their impact on recurrence-free survival defined as the time from randomisation until the earliest occurrence of muscle invasion, distant metastasis, second primary tumour or death due to malignant disease. Schumacher et al (1994) [102] applied the Cox PH model to study the impact of clinical and biological prognostic factors on recurrence-free survival of the patients in this study. They proposed a multivariable regression model which was based on the categorisation of continuous predictors such as age and number of positive lymph nodes. Their proposed linear model comprises 4 prognostic factors tumour grade, number of positive lymph nodes, progesterone respecter, and hormonal treatment, all as categorical variables.

Sauerbrei and Royston (1999) [94] further studied this data set and proposed prognostic models based on the MFP approach [94]. We only consider one of their proposed models, "model III" from Sauerbrei and Royston (1999) [94]. We applied the measures of predictive ability to both the linear model, proposed by Schumacher et al (1994) [102], and the MFP model, proposed by Sauerbrei and Royston (1999) [94].

8.2.4 Data set 4: prostate cancer

The fourth data set is from a well-known trial in patients with advanced prostate cancer. The data set with 506 patients has been analysed by Byar and Green (1980) [14] and others; the data may be found in Reference [7]. Missing values in 31 observations were replaced with imputations to give complete data for analysis of results from all 506 patients. We have applied the MFP approach to this data and computed the predictive ability measures for the resulting model. The MFP model is comprised of 5 continuous prognostic factors: age; standardised weight; acid phosphates; haemoglobin (g=100 ml); and size of primary tumour; and 2 binary prognostic factors - performance status and history of cardiovascular disease. In this model, only acid phosphates is subject to FP1 transformation with power 0.

8.2.5 Data set 5: renal cancer I

The fifth data set is from MRC RE01 randomised trial comparing interferon-α with medroxyprogesterone acetate (MPA) in patients with metastatic renal carcinoma. We analysed data from 347 patients that participated in this randomised trial. The data set consists of clinical and biological prognostic factors of the patients. Missing values were replaced with imputations to give complete data for analysis of results from all patients.

Ritchie et al (1999) [84] studied the effect of two treatments, interferon-α with medroxyprogesterone acetate (MPA), on the overall survival by fitting a multivariable Cox PH model on 335 patients and 236 deaths. The model, with deletion of nonsignificant prognostic factors, resulted in a model comprising WHO performance status, haemoglobin, white cell count and time from metastasis to randomisation. We apply this model to all 347 individuals and compare it with a MFP model in which the variable time from metastasis to randomisation is subject to FP1 transformation with power -0.5.

8.2.6 Data set 6: renal cancer II

The sixth data set uses data from patients with progressive metastatic renal cell carcinoma who were entered into consecutive clinical trials to receive either (A) IFN-α2a, IL-2 (n=102 pts), (B) IFN-α2a, IL-2 and 5-FU (n=235 pts) or (C) IFN-α2a, IL-2 and 5-FU combined with 13cRA (n=88 pts) (Atzpodien et al, 2003 [8]). Patient treatments included radical tumour nephrectomy (n=412), chemotherapy (n=5), immunotherapy

(n=47), chemoimmunotherapy (n=8), and hormone therapy (n=32).

Royston et al (2006) [91] constructed a prognostic model based on 425 of the patients recruited in this study using fractional polynomials considering the overall survival as the outcome. Six binary predictors (sex, lung, lymph node, liver, bone, brain/CNS metastasis) and eight continuous predictors (age, time from diagnosis to metastatic disease, number of metastatic sites, ESR, C-reactive protein (CRP), haemoglobin, neutrophils, LDH) were included in univariate FP analysis. The MFP algorithm selected five prognostic factors as significant at the 5% level in a multivariable model: lymph node metastasis, liver metastasis, bone metastasis, age, CRP, and neutrophils. Royston et al (2006) [91] proposed a model for this data set based on the MFP approach where C-reactive protein was subject to a FP1 transformation with power -2. We used a subset of this data from 322 individuals and applied the model proposed by Royston et al (2006) to compute the measures of predictive ability considering overall survival as outcome. The estimated coefficients in the model and goodness of fit statistics are included in Appendix C.

8.2.7 Data set 7: primary biliary cirrhosis I (PBC I)

The seventh data set is from a study on primary biliary cirrhosis (PBC) which is a degenerative liver disease, often rapidly fatal. In a trial at Mayo Clinic, 312 patients participated in the randomised placebo controlled trial of the drug D-penicillamine. Fleming and Harrington (1991) [27] presented the data in Appendix $D1$ of [27]. The data set contains values on overall survival time (the number of days between randomisation and death), assigned treatment, age, sex, biochemical measurements, and disease conditions.

Lawless (2003) [60] studied this data set and considered 5 covariates as important in predicting survival time. They are: age; oedema, i.e. a variable scaled to take values 0, 0.5, and 1, respectively, denoting three levels of oedema of increasing severity; serum albumin concentration; serum bilirubin concentration; and prothrobin time. Lawless (2003) ([60], page 423) fitted a Cox PH model to this data by considering age and oedema (both untransformed) and log transformation of the last three covariates, i.e. $\ln(albumin)$, $\ln(bilirubin)$, and $\ln(prothrobin)$. We applied the predictive ability measures to this model.

8.2.8 Data set 8: primary biliary cirrhosis II (PBC II)

The eighth data set is from another study on primary biliary cirrhosis (PBC). A total of 248 patients were randomised to receive either azathioprine or placebo (Christensen et al, 1985 [15]). This data set was analysed by Christensen (1985) [15] and later by Royston et al (2006) [90]. After removing 41 (17%) of cases with missing values or no patient follow-up, data on 207 patients (105 deaths) in the PBC data set were available for analysis.

Royston et al (2006) [90] developed a multivariable prognostic model for overall survival using the MFP procedure. They selected variables and functions of continuous variables by using a nominal p-value of 0.05. The Cox model selected by the MFP procedure comprised cirrhosis, central cholestasis, age (untransformed), and log bilirubin. Age, albumin and bilirubin were continuous measurements and the other two were binary. We applied the predictive ability measures discussed in the last three chapters to this model.

8.2.9 Data set 9: lymphoma

The last data set is from a study on diffuse large B cell lymphoma with 240 patients. Rosenwald et al (2007) [85] used this data set to develop a 17-gene classifier of overall survival for patients with advanced diffuse large B cell lymphoma receiving CHOP chemotherapy. A three-level "International Prognostic Index" (IPI) based on both clinical and pathological factors is currently used for risk stratification of patients with aggressive lymphoma (low risk: IPI 0-1, intermediate: IPI 2-3 and high: IPI 4-5). Dunkler et al (2007) [22] evaluated the extent to which the continuous Rosenwald gene score adds to the IPI in the prediction of overall survival in 73 patients of the independent validation series for which the IPI values were available. They computed V_{SchH} for this data set, along with other data sets, to study how effective gene expression profiling is in providing accurate predictions of the survival of individual patients. We computed the measures for two models: model I, including only IPI, and model II, including both IPI and the 17-gene classifier, to evaluate the predictive ability of the proposed gene classifier.

8.3 The estimates of the measures in real data

In this section, the estimates of predictive ability measures computed for the above data sets are presented. Table 8.1 presents the models applied to evaluate the predictive ability

Table 8.1: Summary of the models applied to the data sets, model Chi-squared and degrees of freedom, skewness and kurtosis of the prognostic indices resulting from the fitted resgression models.

Study	Model	Model χ^2	*d.f.*	Sample Size	% Censored	*P.I.* Skewness	*P.I.* Kurtosis
Leg ulcer	MFP I	119.89	5	200	0.52	-2.12	10.21
	MFP II	113.74	5	200	0.52	-5.29	36.73
Breast	RFS I	50.51	8	295	0.60	-0.01	2.46
cancer I	RFS II	72.62	9	295	0.60	-0.16	1.92
	OS I	60.61	8	295	0.73	-0.24	2.19
	OS II	77.64	9	295	0.73	-0.29	1.79
Breast	linear	122.9	5	686	0.56	-0.31	3.07
cancer II	MFP	153.11	6	686	0.56	0.21	3.88
Prostate	MFP	77.41	7	506	0.30	0.40	3.05
Renal I	linear	122.71	6	347	0.07	0.68	6.02
	MFP	132.69	6	347	0.07	0.81	4.96
Renal II	MFP	43.9	1	322	0.15	0.36	2.63
PBC I	Lawless	199.13	5	312	0.60	0.98	3.60
PBC II	Royston	136.81	5	207	0.49	0.31	2.57
Lymphoma	Model I	7.55	2	73	0.34	-0.49	1.32
	Model II	17.64	3	73	0.34	-0.19	2.11

measures. The table contains the χ^2 statistic for each model and the respective degrees of freedom. It also consists of the number of individuals in each data set and censoring proportion. The last two columns are the skewness and kurtosis of the prognostic index (PI), i.e. linear predictor, resulting from the fitted models. Graphs 8-2 to 8-8, presented at the end of this chapter, are histograms of the prognostic index in each model. Tables 8.2 to 8.7 present the estimated predictive ability measures for each model in the different data sets. We discuss the results of each category in the following sections.

8.3.1 Estimates of explained variation measures

Table 8.2 presents the estimates of explained variation measures and the corresponding 95% bootstrap confidence intervals in different studies. From the results of our simulation studies, we expect the measures to agree with each other if the prognostic index of the model is normally distributed with the values of R^2_{OQF} and R^2_{XuOQ} slightly higher and $R^2_{Royston}$ lower. We also expect them to differ as the prognostic index of the model becomes asymmetric. The results of our simulation studies also showed that R^2_D is the only measure that is resistant to the extreme and outlier observations in the data.

Table 8.2 shows that the measures differ substantially if the distribution of the prognostic index is heavily skewed. For example, in both $MFP\ I$ and $MFP\ II$ models fitted

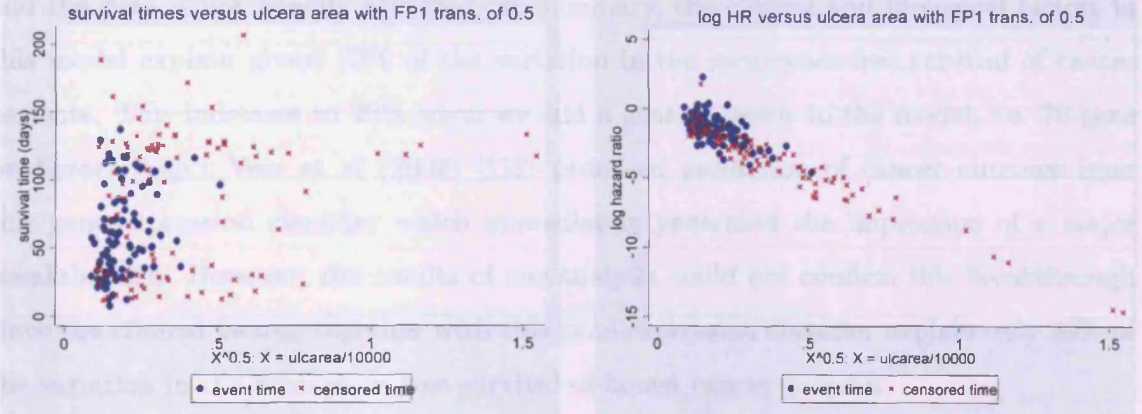
Table 8.2: The estimates of explained variation measures for different studies. The figures in brackets are the bootstrap confidence intervals.

Study	Model	R^2_{PM}	R^2_D	R^2_{OQF}	R^2_{XuOQ}	$R^2_{Royston}$
Leg ulcer	MFP I	0.77	0.54	0.82	0.83	0.60
		(0.64-0.87)	(0.44-0.66)	(0.71-0.91)	(0.71-0.91)	(0.49-0.72)
	MFP II	0.94	0.52	0.92	0.93	0.58
		(0.80-0.98)	(0.44-0.65)	(0.80-0.98)	(0.81-0.98)	(0.46-0.71)
Breast cancer I	RFS I	0.24	0.23	0.34	0.33	0.25
		(0.16-0.39)	(0.16-0.39)	(0.24-0.53)	(0.21-0.51)	(0.17-0.43)
	RFS II	0.32	0.29	0.48	0.44	0.34
		(0.23-0.50)	(0.23-0.46)	(0.36-0.64)	(0.30-0.62)	(0.24-0.53)
	OS I	0.41	0.35	0.54	0.30	0.41
		(0.30-0.67)	(0.25-0.54)	(0.38-0.80)	(0.02-0.77)	(0.30-0.62)
	OS II	0.51	0.41	0.68	0.36	0.50
		(0.40-0.73)	(0.31-0.58)	(0.55-0.87)	(0.07-0.88)	(0.40-0.70)
Breast cancer II	linear	0.24	0.22	0.34	0.35	0.24
		(0.17-0.32)	(0.16-0.29)	(0.26-0.45)	(0.24-0.47)	(0.18-0.32)
	MFP	0.27	0.28	0.37	0.38	0.29
		(0.21-0.35)	(0.21-0.35)	(0.30-0.46)	(0.30-0.48)	(0.23-0.38)
Prostate	MFP	0.13	0.13	0.18	0.16	0.13
		(0.09-0.20)	(0.09-0.21)	(0.13-0.27)	(0.11-0.26)	(0.09-0.19)
Renal I	linear	0.25	0.24	0.34	0.34	0.22
		(0.20-0.35)	(0.19-0.31)	(0.28-0.47)	(0.27-0.46)	(0.18-0.29)
	MFP	0.27	0.26	0.37	0.36	0.24
		(0.21-0.36)	(0.20-0.33)	(0.29-0.46)	(0.30-0.47)	(0.19-0.31)
Renal II	MFP	0.11	0.11	0.14	0.13	0.10
		(0.05-0.18)	(0.05-0.19)	(0.07-0.23)	(0.06-0.22)	(0.04-0.16)
PBC I	Lawless	0.56	0.65	0.69	0.65	0.70
		(0.48-0.65)	(0.55-0.74)	(0.59-0.79)	(0.56-0.76)	(0.60-0.80)
PBC II	Royston	0.58	0.61	0.65	0.63	0.62
		(0.47-0.67)	(0.50-0.70)	(0.56-0.76)	(0.53-0.75)	(0.50-0.74)
Lymph.	Mod. I	0.10	0.09	0.16	0.13	0.09
		(0.02-0.28)	(0.02-0.30)	(0.03-0.42)	(0.02-0.41)	(0.02-0.29)
	Mod. II	0.23	0.23	0.32	0.27	0.21
		(0.11-0.42)	(0.11-0.40)	(0.14-0.59)	(0.08-0.54)	(0.10-0.42)

for leg ulcer data, where skewness=−5.29 and kurtosis=36.73, we observe large variation and unexpectedly high values in the estimates of some measures. Further assessment of this data revealed that there are some extreme observations in one covariate, i.e. "initial ulcer area" which inflate the measures substantially, with the exception of R^2_D. Figure 8-1 consists of two scatter plots of survival time and log hazard ratio versus FP1 transformation of this variable in $MFP\ I$ model. It is evident that there are some censored observations at the extremes of the distribution of this covariate.

To uncover the influence of extreme observations on the measures, we carried out further analysis by removing these observations from the data, a total of 5. Since these observations are censored, the estimated coefficients in the corresponding Cox PH models are almost the same as the models fitted to the complete data, as seen in Appendix C. We refitted both $MFP\ I$ and $MFP\ II$ models to 195 observations; graphs in figure 8-3 show the prognostic indices of the two models. We computed explained variation measures to assess the predictive ability of the two models after removing the extreme observations. Table 8.3 presents the results with the 95% confidence intervals. The results confirm the conclusions of the simulation studies; the measures are in better agreement with R^2_{OQF} and R^2_{XuOQ} higher and $R^2_{Royston}$ lower.

Figure 8-1: Survival time (left) and log hazard ratio (right) versus initial ulcer area with FP1 transformation of 0.5 using model MFP I for leg ulcer data.



For practical purposes, we suggest using a measure from the 5 candidate measures of explained variation to evaluate the predictive ability in each study. Generally, R^2_{XuOQ} can not be recommended since it is not guaranteed to be non-negative. In contrast, the simulation studies showed that this measure is more likely to result in negative values as censoring increases. We also do not recommend $R^2_{Royston}$ due to its poorer performance with regard to the essential properties. Below, we present a measure in each study as a

Table 8.3: The estimates of explained variation measures in the leg ulcer data after removing the censored observations with extreme values.

Measure category	Measure	*MFP I model*	*MFP II model*
Explained variation	R^2_{PM}	0.65 (0.53-0.77)	0.75 (0.59-0.88)
	R^2_D	0.53 (0.40-0.64)	0.50 (0.39-0.63)
	R^2_{OQF}	0.73 (0.62-0.85)	0.78 (0.65-0.90)
	R^2_{XuOQ}	0.73 (0.43-0.74)	0.78 (0.64-0.90)
	$R^2_{Royston}$	0.56 (0.44-0.66)	0.54 (0.42-0.67)

candidate measure of predictive ability.

In both models for the leg ulcer study we recommend R^2_D because there are some extreme observations in the data that are inflating the other measures. It can be concluded that the available prognostic factors explain about 50% of the variation in the outcome.

In breast cancer I study, R^2_{PM} and R^2_D agree in both RFS I, which includes only clinical and biological factors, and RFS II, which includes 70-gene predictor as well. In this study R^2_{OQF} and R^2_{XuOQ} are both higher and $R^2_{Royston}$ lower. For this study we recommend R^2_D for both RFS I and RFS II since the prognostic index of both models is nearly symmetric and the data is not heavily censored. In summary, the clinical and biological factors in this model explain about 23% of the variation in the recurrence-free survival of cancer patients. This increases to 29% when we add a genetic factor to the model, i.e. 70-gene predictor. Van't Veer et al (2002) [112] promised prediction of cancer outcome from this gene-expression classifier which immediately generated the impression of a major breakthrough. However, the results of our analysis could not confirm this breakthrough since the clinical factors together with this gene-expression classifier explain only 29% of the variation in the recurrence-free survival of breast cancer patients.

In a similar study on overall survival of the same patients, we recommend R^2_{PM} for both OS I and OS II models. Royston and Sauerbrei (2004) [93] showed that R^2_D decreases when the covariate or the prognostic index of the model is short tailed. This is due to the effect of short-tailed covariates on the D measure [93]. Therefore, we recommend R^2_{PM} as the candidate measure in this study. It is evident that the 70-gene predictor increases the variation explained in the overall survival, but not substantially. Note the difference in the estimates of R^2_{OQF} and R^2_{XuOQ} for this model, i.e. 0.68 and 0.36 for the OS II. Xu

and O'Quigley (2001) [78] claimed that R^2_{OQF} and R^2_{XuOQ} should be close in practice. The estimates, however, show that they differ substantially in this case. Moreover, the wide bootstrap confidence intervals for R^2_{XuOQ}, $(0.07, 0.88)$, reflects the findings of our simulation studies in chapter 5.

For breast cancer II study, R^2_{PM} and R^2_D agree in both linear and MFP models. However, we recommend R^2_D since it is resistant to outlier and extreme observations. The MFP model has better predictive ability than a linear model where continuous covariates, e.g. age, have been entered in the model as categorical variables. This shows that classifying continuous a prognostic factor into a dichotomy, trichotomy, or more groups diminishes its predictive ability. We can see a similar effect in the renal cancer I study; the predictive ability of MFP model is higher than the linear model.

For lymphoma study, R^2_{PM} and R^2_D agree in both model I and model II. Therefore the three-level "International Prognostic Index" (IPI) explains 10% of the variation in the survival of patients with large-B-cell lymphoma. This increases to only 23% after including the 17-gene classifier to the model. Despite the increase in explained variation, this tells us that much remains to be known about the disease. The difference in estimates of R^2_{OQF} and R^2_{XuOQ}, indicates the inconsistency of R^2_{XuOQ}, specially in model II where the estimates differ noticeably.

As a candidate measure of explained variation, we recommend R^2_D in prostate, renal cancer I, renal cancer II, and PBC II studies. In PBC I study, we recommend R^2_{PM} since the censoring proportion is more than 50% and the prognostic index of the model is close to lognormal distribution. The simulation studies showed that R^2_D increases with the amount of censoring in this case.

8.3.2 Estimates of explained randomness measures

Table 8.4 presents the estimates of explained randomness measures and the corresponding 95% bootstrap confidence intervals in different studies. The table indicates that ρ^2_W, $\rho^2_{W,A}$, ρ^2_{XuOQ}, and ρ^2_k generally agree, with the exception of leg ulcer study. In this study the estimates of ρ^2_W and $\rho^2_{W,A}$ are much higher than those of ρ^2_{XuOQ} and ρ^2_k. This reflects the results of simulations studies presented in table 6.1 which showed that ρ^2_W results in higher values if the covariate or prognostic index of the model is heavily skewed to the left. The presence of extreme observations in the leg ulcer data, as explained above, inflates $\rho^2_{W,A}$ in a similar way to R^2_{PM} since they both depend on the variance of prognostic index

of the model.

To uncover the influence of extreme observations on the explained randomness measures, we carried out further analysis similar to the one on explained variation measures by removing 5 extreme observations in the leg ulcer data set. We refitted both $MFP\ I$ and $MFP\ II$ models to 195 observations and evaluated the measures. Table 8.5 presents the results with the 95% confidence intervals. The results show that all explained randomness measures decrease. However, the decrease in the estimates of ρ^2_W and $\rho^2_{W,A}$ are more noticeable; they decreased about 10% after removing 5 extreme observations. We recommend the estimates of ρ^2_W as the indication of the explained randomness of the models in all data sets except the leg ulcer study. In this study, it is difficult to evaluate the randomness that is explained by the covariate in the model due to the presence of extreme observations.

Table 8.4: The estimates of explained randomness measures for different studies. The figures in brackets are the bootstrap confidence intervals.

Study	Model	ρ^2_W	$\rho^2_{W,A}$	ρ^2_{XuOQ}	ρ^2_k
Leg ulcer	MFP I	0.90	0.84	0.70	0.71
		(0.78-0.95)	(0.73-0.92)	(0.60-0.81)	(0.60-0.81)
	MFP II	0.99	0.96	0.69	0.69
		(0.93-0.99)	(0.85-0.99)	(0.58-0.79)	(0.59-0.79)
Breast cancer I	RFS I	0.33	0.34	0.34	0.35
		(0.24-0.51)	(0.24-0.51)	(0.23-0.52)	(0.24-0.54)
	RFS II	0.42	0.44	0.46	0.46
		(0.33-0.59)	(0.35-0.60)	(0.35-0.63)	(0.36-0.63)
	OS I	0.52	0.53	0.51	0.54
		(0.38-0.74)	(0.41-0.77)	(0.38-0.71)	(0.43-0.73)
	OS II	0.61	0.63	0.62	0.63
		(0.51-0.82)	(0.51-0.82)	(0.52-0.78)	(0.52-0.79)
Breast cancer II	linear	0.34	0.34	0.33	0.34
		(0.26-0.46)	(0.26-0.44)	(0.25-0.43)	(0.26-0.44)
	MFP	0.36	0.38	0.37	0.40
		(0.29-0.47)	(0.30-0.48)	(0.30-0.46)	(0.30-0.48)
Prostate	MFP	0.18	0.20	0.19	0.20
		(0.13-0.27)	(0.14-0.29)	(0.13-0.29)	(0.14-0.29)
Renal I	linear	0.33	0.36	0.31	0.32
		(0.26-0.44)	(0.28-0.46)	(0.25-0.40)	(0.26-0.40)
	MFP	0.33	0.37	0.33	0.34
		(0.27-0.42)	(0.30-0.48)	(0.26-0.42)	(0.28-0.42)
Renal II	MFP	0.16	0.17	0.15	0.15
		(0.08-0.25)	(0.08-0.27)	(0.07-0.24)	(0.07-0.24)
PBC I	Lawless	0.60	0.68	0.71	0.80
		(0.53-0.68)	(0.60-0.76)	(0.60-0.80)	(0.71-0.87)
PBC II	Royston	0.65	0.69	0.76	0.73
		(0.56-0.74)	(0.60-0.78)	(0.59-0.79)	(0.63-0.82)
Lymph.	Mod. I	0.15	0.15	0.14	0.15
		(0.04-0.37)	(0.03-0.40)	(0.03-0.37)	(0.03-0.38)
	Mod. II	0.32	0.33	0.30	0.31
		(0.15-0.53)	(0.15-0.54)	(0.14-0.52)	(0.15-0.54)

Table 8.5: The estimates of explained randomness measures in the leg ulcer data after removing the censored observations with extreme values.

Measure category	Measure	*MFP I model*	*MFP II model*
Explained variation	ρ^2_W	0.79	0.0.91
		(0.69-0.88)	(0.78-0.97)
	$\rho^2_{W,A}$	0.77	0.86
		(0.66-0.85)	(0.72-0.95)
	ρ^2_{XuOQ}	0.67	0.66
		(0.56-0.79)	(0.54-0.77)
	ρ^2_k	0.69	0.66
		(0.57-0.79)	(0.55-0.78)

Table 8.6: The three time points (in days) at which the predictive ability of the models are evaluated using the Graf et al's measure (1999) for each study.

Study	Three time points		
	T_1	T_2	T_3
Leg ulcer	33	54	86
Breast cancer I (RFS)	618	1078	1986
Breast cancer I (OS)	931	1386	2109
Breast cancer II	426	646	1105
Prostate	289	715	1142
Renal I	85	223	455
Renal II	291	554	1061
PBC I	597	1083	2071
PBC II	456	1024	1744
Lymphoma	201	420	1114

8.3.3 Estimates of predictive accuracy measures and R^2_{SchK}

In this section, we present the estimates of predictive accuracy measures and R^2_{SchK} proposed by Schemper and Kaider (1997). We evaluate the measure proposed by Graf et al (1999), $R^2_G(T)$ at three time points in each study. The time points are the $0.25th$, $0.50th$, and $0.75th$ quantile of the time to the last event in each study. Table 8.6 displays the time points, in days, in each study. For practical purposes, it is worth bearing in mind that the choice of time point will be application-specific. In some studies, there might be clinically relevant fixed time point, such as five-year survival being used as the effectiveness of a specific treatment, e.g. chemotherapy.

Table 8.7 shows the estimated values and the 95% bootstrap confidence intervals of the measures. The estimates of predictive accuracy measures $R^2_G(T^*)$ and V_{SchH} are lower than the corresponding explained variation and explained randomness measures. The estimates are the highest in the leg ulcer and PBC I and II studies with about 40%. The estimates of predictive accuracy measures in most of the other studies are below 20%. This indicates the limited ability of the available clinical and biological prognostic factors in predicting the individual status of patients in terms of experiencing the event of interest by means of the Cox PH regression model.

As it was observed in the simulation studies, the estimates of $R^2_G(T^*)$ increases with increasing T^* in most of the studies. The estimates of $R^2_G(T^*)$ and V_{SchH} are fairly close for the fitted models in each study which indicates that the gain in terms of predictive accuracy is very limited even after using a more representative model, i.e. models based on MFP approach. The estimates of R^2_{SchK} are much higher than the predictive accuracy

measures. The results indicate that this measure is in agreement with the explained variation measures studied in chapter 5, as shown in the simulation studies.

The estimates of $R^2_G(T^*)$ and V_{SchH} for the models for the breast cancer I and lymphoma studies confirm our previous findings that the gene-expression classifiers, i.e. 70-gene predictor in breast cancer and 17-gene predictor in lymphoma, have limited predictive ability. For example, the estimates of $R^2_G(T_3)$ and V_{SchH} for the OS I model in breast cancer study, i.e. model with only clinical prognostic factors, are 0.20. They increase to 0.23 and 0.24 when we add the gene-expression classifier to the model. The predictive accuracy of both models increased, but not substantially. This contradicts the conclusion made by Van't Veer et al (2002) [112] that this gene-expression classifier is strongly predictive of the survival of breast cancer patients characterised in this study.

Table 8.7: The estimates of predictive accuracy measures and Schemper and Kaider measure (1997) for different studies. The figures in brackets are the bootstrap confidence intervals.

Study	Model	Graf measure at three time points			V_{SchH}	R^2_{SchK}
		$R^2_G(T_1)$	$R^2_G(T_2)$	$R^2_G(T_3)$		
Leg ulcer	MFP I	0.23	0.33	0.46	0.39	0.62
		(0.08-0.37)	(0.18-0.42)	(0.33-0.57)	(0.30-0.47)	(0.50-0.73)
	MFP II	0.22	0.32	0.46	0.38	0.62
		(0.08-0.33)	(0.19-0.43)	(0.34-0.57)	(0.30-0.46)	(0.50-0.73)
Breast cancer I	RFS I	0.12	0.12	0.17	0.14	0.29
		(0.04-0.17)	(0.04-0.19)	(0.09-0.24)	(0.10-0.3)	(0.18-0.44)
	RFS II	0.12	0.14	0.23	0.19	0.39
		(0.04-0.18)	(0.06-0.21)	(0.14-0.30)	(0.15-0.28)	(0.30-0.52)
	OS I	0.13	0.17	0.20	0.20	0.46
		(0.03-0.21)	(0.07-0.26)	(0.10-0.28)	(0.14-0.31)	(0.32-0.63)
	OS II	0.11	0.17	0.23	0.24	0.54
		(0.02-0.18)	(0.07-0.25)	(0.12-0.31)	(0.18-0.37)	(0.45-0.70)
Breast cancer II	linear	0.09	0.13	0.17	0.16	0.28
		(0.05-0.13)	(0.08-0.18)	(0.11-0.22)	(0.12-0.20)	(0.22-0.35)
	MFP	0.12	0.16	0.20	0.18	0.30
		(0.07-0.18)	(0.10-0.21)	(0.14-0.25)	(0.14-0.23)	(0.24-0.38)
Prostate	MFP	0.06	0.11	0.10	0.10	0.15
		(0.02-0.10)	(0.06-0.15)	(0.05-0.14)	(0.07-0.14)	(0.10-0.24)
Renal I	linear	0.21	0.27	0.18	0.19	0.35
		(0.13-0.28)	(0.20-0.33)	(0.10-0.24)	(0.15-0.24)	(0.29-0.45)
	MFP	0.24	0.27	0.19	0.20	0.37
		(0.16-0.31)	(0.21-0.34)	(0.11-0.26)	(0.16-0.24)	(0.30-0.46)
Renal II	MFP	0.11	0.13	0.05	0.09	0.15
		(0.06-0.16)	(0.08-0.19)	(0.01-0.12)	(0.05-0.13)	(0.09-0.24)
PBC I	Lawless	0.38	0.47	0.47	0.40	0.54
		(0.19-0.52)	(0.38-0.58)	(0.34-0.57)	(0.34-0.48)	(0.47-0.64)
PBC II	Royston	0.34	0.35	0.43	0.41	0.61
		(0.16-0.49)	(0.20-0.47)	(0.38-0.55)	(0.33-0.49)	(0.51-0.70)
Lymph.	Mod. I	0.05	0.11	0.09	0.08	0.15
		(0.01-0.10)	(0.01-0.18)	(0.06-0.20)	(0.01-0.19)	(0.02-0.36)
	Mod. II	0.16	0.22	0.24	0.17	0.31
		(0.02-0.24)	(0.05-0.34)	(0.07-0.38)	(0.09-0.34)	(0.16-0.53)

8.4 Discussion

In this chapter, we applied the predictive ability measures studied in chapters 5 to 7 to real data sets from different diseases. Our main objective was to illustrate the application of these measures in medical research. It is important for medical investigators to realise that even strong and highly significant regression coefficients associated with prognostic factors of outcome may not automatically translate into sufficiently accurate prediction or close determination of individual outcome values of patients. Gains from the use of prognostic factors can only be demonstrated by the use of a suitable measure of predictive ability, but not by means of large hazard ratios, nor by their corresponding p-values. This issue often is not taken into account and even partly explains why so many identified prognostic factors "fail" particularly when used to predict outcomes for individual patients.

Furthermore, we have shown how to study the clinical importance of new genetic factors in addition to clinical characteristics of the patients. The results suggest that determining the patients outcome is very limited even after considering the gene classifier in breast cancer I study. As mentioned before, the early papers promising prediction of cancer outcome from this gene classifier generated the impression of a major breakthrough. Our results could not confirm such a breakthrough.

Our second objective was to compare the measures with real data sets and provide justification for the observed discrepancies in the estimates of measures. We compared the measures and recommended a measure for practical applications in each study. The measures within explained variation and explained randomness groups are broadly in agreement if the distribution of the prognostic index of the model is approximately normal.

Table 8.8 presents the skewness and kurtosis of prognostic indices of the fitted models together with the range of estimated values in explained variation and explained randomness categories. It is evident that the measures in each class result in similar values if the skewness and kurtosis of the prognostic index of the model is close to that of normal distribution (i.e. skewness=0 and kurtosis=3). The estimated values in each category differ substantially when the skewness and kurtosis of the prognostic index of the model is far from normality. Finally, the limitations of the proposed measures make it impossible to recommend one measure for all the studies.

By applying the measures to linear, where the continuous covariates transformed to categorical variable, and MPF models, we showed the application of the measures in sta-

Table 8.8: The range of explained variation and explained randomness estimates for each study.

Study	Model	*P.I.* Skewness	*P.I.* Kurtosis	Range in *E.V.* measures	Range in *E.R.* measures
Leg ulcer	MFP I	-2.12	10.21	0.29	0.20
	MFP II	-5.29	36.73	0.42	0.30
Breast cancer I	RFS I	-0.01	2.46	0.10	0.02
	RFS II	-0.16	1.92	0.19	0.04
	OS I	-0.24	2.19	0.24	0.02
	OS II	-0.29	1.79	0.32	0.02
Breast cancer II	linear	-0.31	3.07	0.13	0.01
	MFP	0.21	3.88	0.11	0.04
Prostate	MFP	0.40	3.05	0.05	0.02
Renal I	linear	0.68	6.02	0.12	0.05
	MFP	0.81	4.96	0.13	0.04
Renal II	MFP	0.36	2.63	0.04	0.02
PBC I	Lawless	0.98	3.60	0.14	0.20
PBC II	Royston	0.31	2.57	0.07	0.11
Lymphoma	Model I	-0.49	1.32	0.07	0.01
	Model II	-0.19	2.11	0.11	0.02

tistical practice. The results showed that the models developed using MFP approach have better predictive ability. Therefore, as Royston et al (2006) [90] indicated, dichotomising continuous covariates diminishes the overall predictive ability of the models.

Figure 8-2: Distributions of the prognostic index in the MFP I (left) and MFP II (right) models for leg ulcer study.

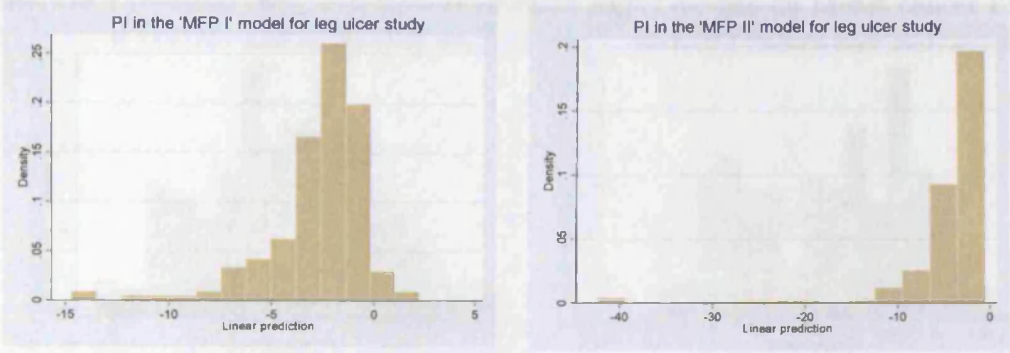


Figure 8-3: Distributions of the prognostic index in the MFP I (left) and MFP II (right) models for leg ulcer study after removing the censored observations with extreme covariate values.

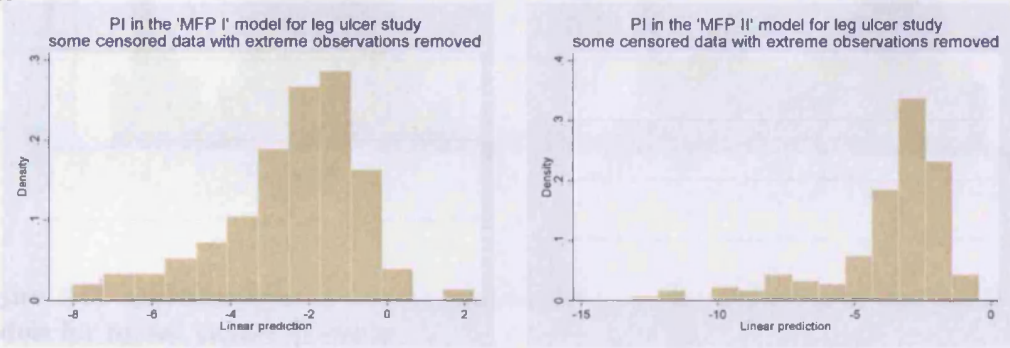


Figure 8-4: Distributions of the prognostic index in the RFS I (top left), RFS II (top right), OS I (bottom left), and RFS II (bottom right) models for breast cancer I study.

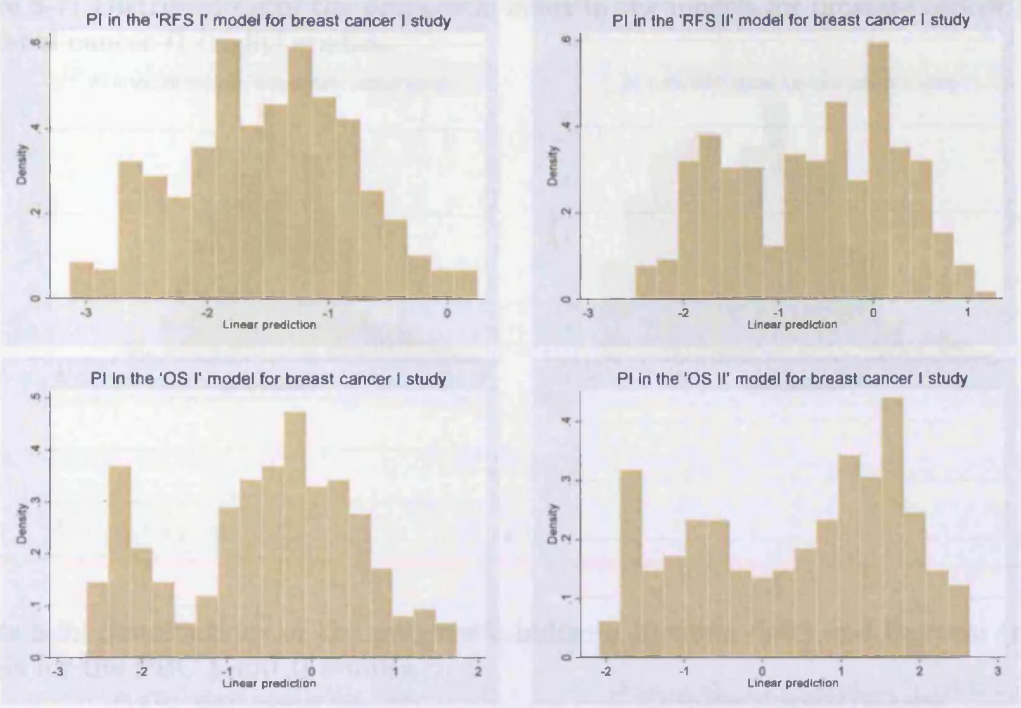


Figure 8-5: Distributions of the prognostic index in the linear (left) and MFP (right) models for breast cancer II study.

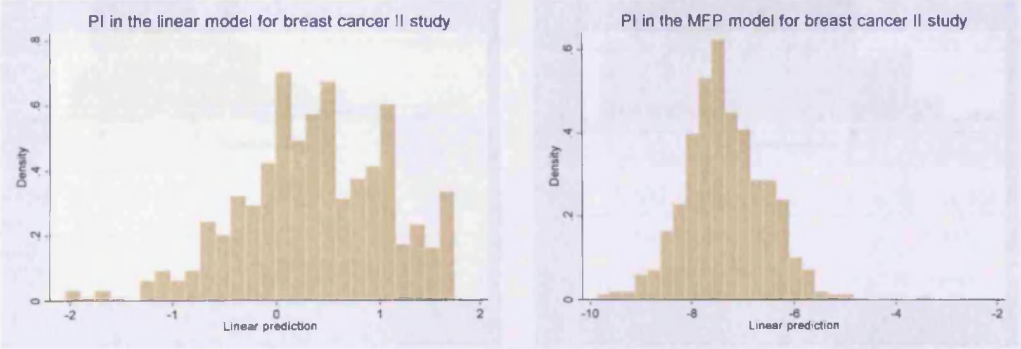


Figure 8-6: Distributions of the prognostic index in the linear (left) and MFP (right) models for renal cancer I study.

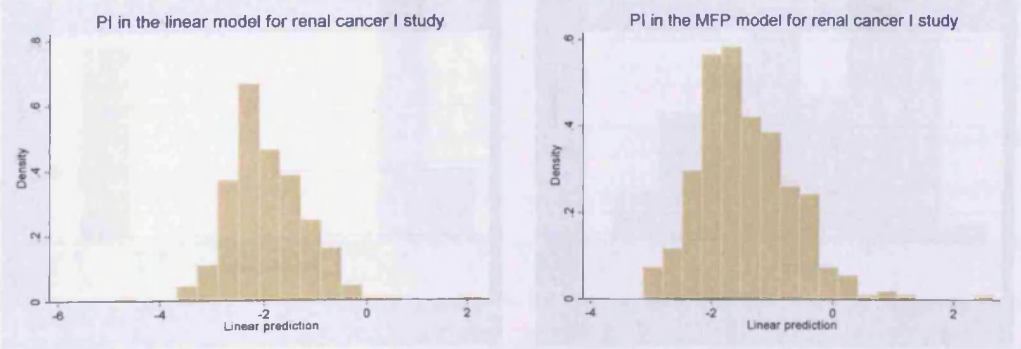


Figure 8-7: Distributions of the prognostic index in the models for prostate cancer (left) and renal cancer II (right) studies.

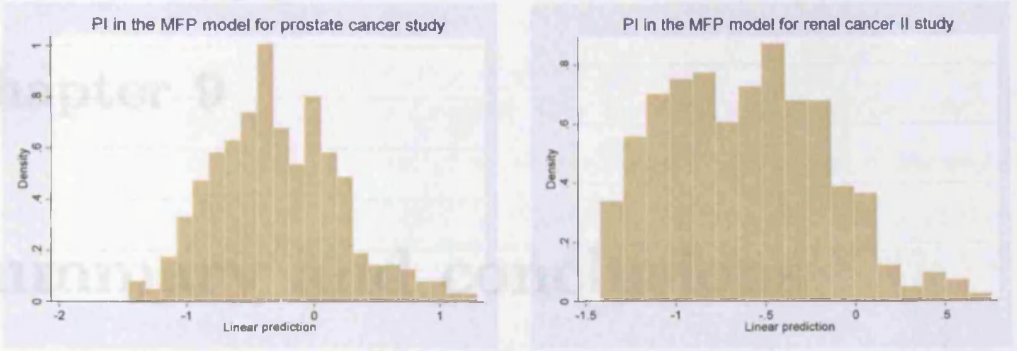


Figure 8-8: Distributions of the prognostic index in Fleming (left) and Royston (right) models for the PBC I and II studies.

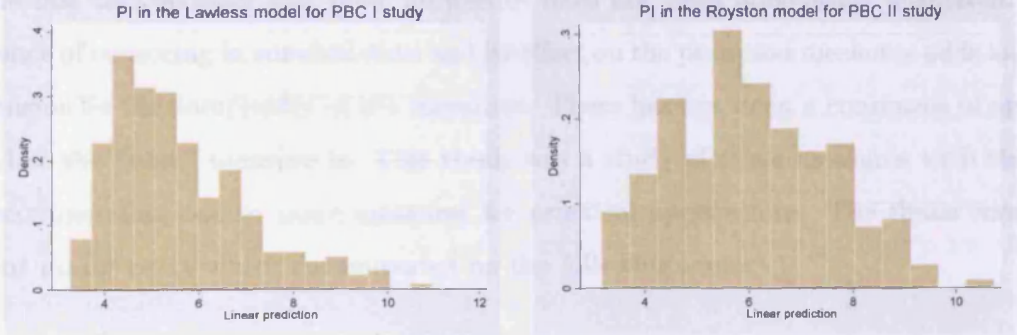
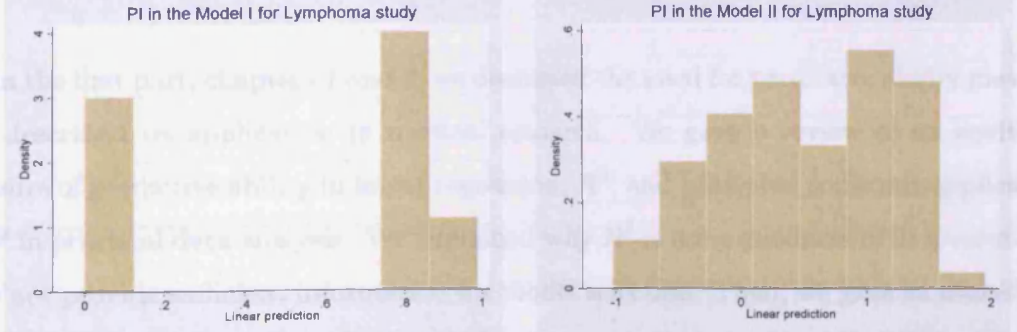


Figure 8-9: Distributions of the prognostic index in the model I (left) and model II (right) for the lymphoma study.



Chapter 9

Summary and conclusions

9.1 Summary

Several measures have been proposed to evaluate the predictive ability of survival models in the last two decades and their properties have not been adequately evaluated. The presence of censoring in survival data and its effect on the proposed measures adds another dimension to the complexity of the measures. There has not been a consensus of opinion on what the "best" measure is. This thesis was a study of these measures with the aim of recommending one or more measures for practical applications. The thesis consisted of four major parts which concentrated on the following issues:

1) the need for predictive ability measures and the measures for survival models

2) critical review of the proposed measures and the need for further research

3) presenting the results of simulation studies on the proposed measures

4) applications of the measures in prognostic modelling.

In the first part, chapters 1 and 2, we discussed the need for predictive ability measures and described its application in medical research. We gave a review of an equivalent measure of predictive ability in linear regression, R^2, and presented some mis-applications of R^2 in practical data analysis. We explained why R^2 is not a goodness of fit measure and does not provide sufficient information for model selection. Then, we gave an overview of the proposed measures of predictive ability for survival models, mainly for the Cox PH model, by classifying them into three main categories:

I) explained variation measures

II) explained randomness measures

III) predictive accuracy measures.

The classification of proposed measures into three major groups has been a main theme of this thesis. The theoretical underpinning and conceptual differences of the proposed measures has led us to this classification. We have also identified two other measures that could not be classified into one of the three main classes. These comprised an "other" category.

In the second part, chapters 3 and 4, we provided the framework for examining the proposed measures. To study the measures systematically, we defined two sets of properties, i.e. essential and desirable, that a measure of predictive ability should possess in the context of survival analysis. Chapter 3 described the essential and desirable properties of a suitable measure of predictive ability for survival models. Some of the criteria are based on or closely related to those proposed by Schemper and Stare (1996) [99] and Royston and Sauerbrei (2004) [93] for a "good" measure of explained variation. The essential properties of a suitable measure are independence of censoring, independence of sample size, and parameter and number of variables monotonicity. In our opinion, these are the properties that a measure of predictive ability should possess in the context of survival analysis. The desirable properties include robustness, generalisability, the availability of straightforward confidence intervals, and partial and adjusted measures. We then considered the measures which have been proposed against the essential criteria. The shortcomings of some measures with respect to essential properties led us to a short-list, called potentially recommendable measures, requiring further investigation of properties. From a total of 10 potentially recommendable measures, 5 were classified in the explained variation category, 3 in the explained randomness category, and 2 in the predictive accuracy category. We also included one measure from the "other" category in our investigations because it was potentially recommendable.

In chapter 4, we set out the need for further investigation of the potentially recommendable measures and proposed comprehensive simulation studies to explore the measures further. The rest of chapter 4 presented the simulation design and different parameters involved in the simulation studies. Mostly, the simulation studies were univariate in character. In the simulation studies, we considered 4 covariate distributions

with different skewness to model potentially different distributions of the prognostic index of the multiple regression model. We put more emphasis on the normally distributed covariate because by virtue of the central limit theorem the prognostic index of a multiple regression model, which is usually a function of several random variables, tends to Normality as the dimension of the parameter vector β increases. In the next section, we will discuss the findings of the simulation studies.

9.2 Findings of the simulation studies

In the third part of this thesis, chapters 5 to 7, we presented the results of simulation studies on three classes of predictive ability measures, i.e. explained variation measures, explained randomness measures, predictive accuracy measures, and the measure proposed by Schemper and Kaider (1997) [98], R^2_{SchK}. The performance of the measures with respect to the criteria outlined in chapter 3 is summarised below.

9.2.1 Explained variation measures

In chapter 5, we carried out simulation studies on 5 potentially recommendable measures in the explained variation category. The measures proposed by Korn and Simon (1990) [53] and Akazawa (1997) [2] were excluded from our studies because previous simulation studies provided us with sufficient evidence that these measures are affected by the amount of censoring.

The results of the simulation studies, presented in chapter 5, show that the explained variation measures are influenced by the distribution of covariate or prognostic index in the case of multiple regression, with the exception of R^2_{PM}. The results also indicate that R^2_{XuOQ} (Xu and O'Quigley (2001) [78]) and $R^2_{Royston}$ (Royston (2006) [88]) perform poorly with respect to censoring. The measure proposed by Xu and O'Quigley (2001), R^2_{XuOQ}, can not be guaranteed to be non-negative, and $R^2_{Royston}$ is heavily influenced by the degree of censoring. Therefore, we reject these two measure. The measure proposed by O'Quigley and Flandre (1994), R^2_{OQF}, is slightly affected by the amount of censoring but performs reasonably well in general with respect to the other essential properties. Both R^2_{OQF} and R^2_{XuOQ} possess parameter and variable monotonicity properties. But, R^2_{XuOQ} performs poorly in censored conditions since the chance that it decreases after adding a new independent covariate to the model is more than for the other measures.

Moreover, R^2_{OQF} and its modification, R^2_{XuOQ}, quantify the variation in the covariate which is explained by the survival time. This makes both measures, R^2_{OQF} and R^2_{XuOQ}, counter-intuitive.

The results of the simulation studies indicate that two measures R^2_{PM} (Helland (1987) [41] and Kent & O'Quigley (1988) [49]) and R^2_D (Royston & Sauerbrei (2004) [93]) satisfy the essential criteria. The results of our simulation studies, however, revealed some limitations of R^2_{PM} and R^2_D with respect to the desirable properties. The results indicate that R^2_{PM} is influenced by the presence of extreme or outlier observations since it depends on the variance of the prognostic index of the model. R^2_D is not influenced by extreme and outlier observations, but is affected by the covariate distribution. Therefore, if the distribution of one (or some) prognostic factor(s) in a study is either contaminated with outliers or skewed, the value of R^2_{PM} and R^2_D may be considerably different from what would have been achieved had the outlier contamination or skewness not been present. Furthermore, our simulation results showed that all of the explained variation measures increase with the amount of censoring in under-fitted models.

9.2.2 Explained randomness measures

In chapter 6, we presented the results of simulation studies on 3 potentially recommendable measures in explained randomness category. Explained randomness measures comprise an alternative class of measures. These measures are essentially founded on the concept of information, and the way information is quantified in communication theory (Shannon (1948) [104]). Kullback and Leibler (1951) [55] applied this concept to statistics and established the relationship between information gain [55] and R^2 in linear regression.

The shortcomings of explained randomness measures proposed by Nagelkerke (1991) [71], Magee (1990) [68], Maddala (1983) [67], and Verweij and Van Houwelingen (1993) [113] with respect to essential properties has led us to the 3 potentially recommendable measures in the explained randomness category, proposed by Kent and O'Quigley (1988) [49], Xu and O'Quigley (1999) [116], and O'Quigley et al (2005) [80].

The results of the simulation studies, presented in chapter 6, show that the explained randomness measures, i.e. ρ^2_W, ρ^2_{XuOQ}, and ρ^2_k, are influenced by the distribution of covariate or prognostic index in the case of multiple regression. The results indicate that among the randomness measures, ρ^2_k (O'Quigley et al (2005) [80]) performs the worst with regard to censoring and because of this is not recommended as a candidate measure of

explained randomness. The measure proposed by Xu and O'Quigley (1999) [116], ρ^2_{XuOQ}, performs well in random censoring conditions.

The measure proposed by Kent and O'Quigley (1988) [49], ρ^2_W, and its approximation $\rho^2_{W,A}$ are almost unaffected by the amount of censoring and generally satisfy the essential properties. The results also show that $\rho^2_{W,A}$ is not a good approximation to ρ^2_W if the covariate distribution is asymmetric. Two measures of explained randomness, ρ^2_W and ρ^2_{XuOQ}, performed well with respect to the essential properties; but, ρ^2_{XuOQ} is not straightforward to interpret since it evaluates the randomness in the covariate, which is explained by survival time.

The results of further simulation studies indicate limitations in all of the explained randomness measures similar to those of R^2_{PM}. They are all affected by extreme and outlier observations. Also, the measures increase with the amount of censoring in underfitted models.

9.2.3 Predictive accuracy measures & R^2_{SchK}

In chapter 7, we presented the results of simulation studies on predictive accuracy measures and the measure proposed by Schemper and Kaider (1997) [98], R^2_{SchK}. We excluded Schemper's V_1 and V_2 measures (1990) (1994) [95] [96] from our studies because previous studies showed that these two measures are influenced to a major extent by the amount of censoring.

The results in chapter 7 indicate that R^2_G (Graf et al (1999) [31]), V_{SchH} (Schemper and Henderson (2000) [97]), and R^2_{SchK} (Schemper and Kaider (1997) [98]) are affected by the covariate distribution. The results show that R^2_G and R^2_{SchK} perform well in both random and type I censoring. Also, V_{SchH} is not affected by random censoring if the covariate is normally distributed or skewed to the left, whereas it is affected by type I censoring. All three measures possess parameter monotonicity properties. Among the three measures, however, R^2_G performs the worst with regard to the number of variables monotonicity since the chance that it decreases after adding new independent covariate to the model is more than for the other measures. All three are sensitive to outliers and extreme observations in the data. Moreover, V_{SchH} and R^2_{SchK} increase with the amount of censoring if the covariate in mis-modelled, whereas the expected value of R^2_G does not change with increasing censoring.

9.2.4 Comparison of three groups of measures

In summary, the results of our simulation studies have revealed that the expected values of explained randomness measures are higher than the corresponding values of explained variation and predictive accuracy measures. They have also shown that predictive accuracy measures result in lower values than the measures in the other categories. The expected value of R^2_{SchK} agrees with the corresponding value of explained variation measures. All the measures increase with increasing covariate effect and appear to have an upper bound of less than 1. Predictive accuracy measures, however, reach high values, i.e. more than 0.80, only if the covariate effect is unrealistically high. The sampling distribution of all measures shows considerable skewness when censoring is more than 50%. We have also learned that for all the measures, when there is a weak association between the covariate and the outcome and the amount of censoring is high, the sample estimator has a positive bias.

Finally, we update table 3.1 of chapter 3 after our investigation and present it in table 9.1. It is evident from this table that our investigation has led us to reach new conclusions about the properties of some measures. For example, based on previous investigations, in table 3.1 we concluded that Xu & O'Quigley (1999) measure [116], ρ^2_{XuOQ}, was independent of censoring. However, our investigations showed that this measure is affected by type I or administrative censoring. Regarding the desirable properties of the potentially recommendable measures presented in table 3.2, we only carried out investigation on the robustness property of these measures. Thus we have not updated this table in this section.

Table 9.1: Summary of the essential properties of potentially recommendable measures of predictive ability in survival analysis after our investigation

Measure Category	Measure	Proposed by	I	II	III a	III b
Explained Variation	R^2_{PM}	Helland; Kent & O'Quigley (1988)	yes	yes[5]	yes	yes
	R^2_{OQF}	O'Quigley & Flandre (1994)	no	yes[5]	yes	yes
	R^2_{XuOQ}	Xu & O'Quigley (2001)	yes[2]	yes[5]	yes	yes
	R^2_D	Royston & Sauerbrei (2004)	no[3]	yes[5]	yes	yes
	$R^2_{Royston}$	Royston (2006)	no	yes[5]	yes	yes
Explained Randomness	ρ^2_W	Kent & O'Quigley (1988)	yes	yes[5]	yes	yes
	ρ^2_{XuOQ}	Xu & O'Quigley (1999)	no[4]	yes[5]	yes	yes
	ρ^2_k	O'Quigley et al (2005)	no	yes[5]	yes	yes
Predictive Accuracy	R^2_G	Graf et al (1999)	yes	yes[5]	yes	yes
	V_{SchH}	Schemper & Henderson (2000)	no[1]	yes[5]	yes	yes
Other	R^2_{SchK}	Schemper & Kaider (1997)	yes	yes[5]	yes	yes

Key of the table

I) Independence of censoring; II) Independence of sample size

III-a) Parameter monotonicity III-b) number of variables monotonicity

yes: the measure does possess the desired property

no: the measure does not possess the desired property

1) This measure is largely independent of random censoring if the covariate is normally distributed or skewed to the left

2) The expected value of this measure does not change with censoring but results in negative values

3) This measure is independent of censoring if the covariate is normally distributed

4) The expected value of this measure does not change with censoring in random censoring conditions with normal covariates

5) Sample size has a moderate effect, i.e. positive bias, when there is a weak association and censoring is high

9.3 Applications of the measures in medical research

In the last part of this thesis, chapter 8, we applied the measures to data sets from different disease types to quantify the predictive ability of available/known prognostic factors. We applied the measures to different models and discussed the observed discrepancies in the estimated values of the measures based on the results of simulation studies. Two important findings resulted. First, the measures within each category are broadly in agreement if the distribution of the prognostic index of the model is approximately normal. Second, the estimated values of R^2_{OQF} and its modification R^2_{XuOQ} in the study of overall survival for breast cancer study I differ substantially in heavily censored data. The results of simulation studies on R^2_{XuOQ} in chapter 5 indicated that this measure behaves inconsistently in heavily censored data; the probability that it results in negative values increases with increasing censoring.

In summary, the results of our analysis on real data sets indicate that the measures within each category differ substantially when the censoring is high or the distribution of the prognostic index of the model is far from normality.

9.4 Recommendations for practice

One of the aims in this study was to recommend a small number of measures for general use. We have classified the measures into three main categories. This classification is a broad but conceptual one. In this section, we first summarise the conceptual differences of the three classes of measures. We then suggest two measures in the explained variation category for general use.

In practice, an important question might be raised: which class of measures should one use to quantify the predictive ability in survival models? The choice of the measure depends on the clinical aim of the study. Two quite different goals can be sought in clinical research. These are the goal of understanding and the goal of prediction. Theoretically, explained variation measures are used if the goal is understanding and predictive accuracy measures are used if the goal is prediction. However, the performance of some measures might make them less useful in practical applications.

Explained variation measures generally quantify how much of the variation in the outcome variable is explained by the predictors in the model. Predictive accuracy measures

evaluate the predictions made in terms of model-based survival probabilities, with and without covariates, and compare them with the survival status of individuals at time t^*. This leads to a measure which shows the relative gain in terms of the accuracy of estimated survival probability in predicting the individuals' status as "dead" or "alive" when using prognostic factors information compared with when not using them. Explained variation measures can be used to quantify the clinical significance of the prognostic factors in the model, whereas predictive accuracy measures can help researchers where they need to know the ability of the prognostic factors in predicting an individuals' status, for example, 2 years after the start of study. Explained variation measures are intuitive and easy to explain to researchers in medical research, whereas it is more difficult to interpret the estimates of R^2_G or V_{SchH}.

Explained randomness comprises an alternative class of measures. These measures are founded on the way information is quantified in communication theory (Shannon (1948) [104]). Kullback and Leibler (1951) [55] applied this concept to statistics and established the relationship between information gain [55] and R^2 in linear regression. However, the interpretation of these measures is a challenge in models other than linear regression since they generalise the relationship between the information gain and R^2, presented by Kullback and Leibler (1951) [55]. Nevertheless, they can be interpreted as the information in the outcome, as defined in information theory, which can be potentially recovered by the prognostic factors in the model.

In the next sections we present our recommended measures and provide justification for the recommendations.

9.4.1 Explained variation measures - recommended

We recommend explained variation measures in general and in particular R^2_{PM} and R^2_D for general use. First, they are interpretable and easy to explain to clinicians compared with measures in the other two groups. For example, an estimate of 0.20 for R^2_{PM} means that 20% of the variation in the outcome is explained by the prognostic factors in the model. It is more difficult to interpret the same estimate of explained randomness or predictive accuracy measures.

Second, they offered good performance in our studies and mostly satisfied the essential criteria defined in chapter 3. The amount of censoring and censoring mechanism do not affect R^2_{PM}. This also applies to R^2_D if the distribution of the prognostic index of the

model is not skewed. Other measures are either influenced by the censoring mechanism or affected by the follow-up period, with the exception of ρ^2_W and R^2_{SchK}.

Third, R^2_{PM} and R^2_D have traceable statistical properties and can be consistently estimated, whereas the statistical properties of explained randomness measures and predictive accuracy measures are difficult to establish, especially in the context of multiple regression models.

Fourth, the estimates of R^2_{PM} and R^2_D appear to give a good reflection of strength of association as measured by the covariate effect, β, and tend to 1 for high, but plausible, values of β.

Fifth, R^2_{PM} and R^2_D are based on the same principle thus can be used for sensitivity analysis. We recommend computing both R^2_{PM} and R^2_D for any study. If there is a large discrepancy between them, say more than 10%, we suggest investigating the data under study for the potential reasons such as the presence of outlier observations or highly skewed prognostic index. Our simulation studies indicate that R^2_D can not be larger than R^2_{PM} except in heavily censored data where the prognostic index of the model is highly skewed, or in cases where the data contains some outlier observations which affect R^2_{PM}.

Sixth, both R^2_{PM} and R^2_D can be generalised for use in the flexible parametric models proposed by Royston and Parmar (2002) [92].

Finally, R^2_D has the advantage that it can be used in a model validation context which is an important part of prognostic modelling.

9.4.2 Explained randomness measures - not recommended

We do not recommend explained randomness measures for the following reasons. First, as explained in the last section, the proposed explained randomness measures use the relationship between the correlation coefficient of two normally distributed random variables and Kullback-Leibler information gain [55] to define the proposed explained randomness measures for survival models. Despite all the promising properties of these measures, the explained randomness measures lack clear interpretation.

Second, the results of our investigation indicate that the measure proposed by Kent and O'Quigley (1988), ρ^2_W, is the only measure which has performed satisfactorily with respect to the essential criteria. In this measure, the baseline hazard in the Cox PH model is replaced with a specific function of time to form the Kullback-Leibler information gain

[55]. The procedure, however, is not straightforward and inference for the resulting estimate is even less so (O'Quigley (2008) [74]).

Third, ρ^2_W is complex to calculate. Although a very simple approximation, $\rho^2_{W,A}$, was suggested, our simulations studies showed that $\rho^2_{W,A}$ is not a good approximation if the prognostic index of the model is non-normal.

Fourth, the measures in this category lack generalisability. The measures ρ^2_{XuOQ} and ρ^2_k accommodate time-dependent covariates in the context of the Cox PH model. However, all three measures ρ^2_W, ρ^2_{XuOQ}, and ρ^2_k are based on specific properties of the Cox PH model which makes them difficult, if not impossible, to be generalised to other types of survival models, e.g. flexible parametric models proposed by Royston and Parmar (2002) [92].

9.4.3 Predictive accuracy measures - not recommended

The main drawback in both predictive accuracy measures, R^2_G and V_{SchH}, is their dependence on the follow-up period. This limits their applications, specially when comparing studies with different follow-up periods.

Furthermore, our simulation studies as well as analysis of real data in chapter 8 indicate that predictive accuracy measures are generally lower than explained variation and explained randomness measures. The lower values in predictive accuracy measures are expected since they capture the uncertainty in a binary outcome, i.e. event status as being "dead" or "alive", accounted for by a model rather capturing the uncertainty about the survival time itself. In other words, at each event time a binary outcome is evaluated which leads to lower values of predictive ability, due to the loss of information. This approach is similar to the R^2 analogues for logistic regression [20] that compare discrete observed values (typically zero and one for a dichotomous dependent variable) with predicted probabilities that result from applying logistic regression.

The only measure in the "other" category proposed by Schemper and Kaider (1997), R^2_{SchK}, performed well with regard to the essential criteria. It is, however, a non-parametric measure of association, numerically complex, and not affording clear interpretation.

9.5 Conclusions and outlook

This thesis has studied the measures of predictive ability proposed for survival models, with a particular emphasis on measures proposed for the Cox PH model. It has explained their use in medical research, and systematically compared their performance with respect to a set of criteria.

As described by the authors of the measures, most of them possess promising properties. They, however, have shortcomings as addressed in this thesis. Therefore, there is not a single measure of predictive ability that can be universally recommended. Nonetheless, findings from our studies present a good though not unassailable case for preferring the explained variation category in general and two of the measures specifically, i.e. R^2_{PM} and R^2_D, over the other predictive ability measures applicable to survival models.

Finally, we have summarised the conclusions of our studies in some flow diagrams which can be used as a guide to choose the right measure. The flow diagram in figure 9-1 guides users to choose the right explained variation measure(s). No measure is recommended when the prognostic index of the model is asymmetric and outliers are present. This condition is highlighted with a question mark in the diagram. This can be an area for further research. We have also prepared similar diagrams for the potentially recommendable measures in the explained randomness and predictive accuracy category (Figures 9-2 and 9-3). They can be used as a guide in choosing the right measure if one wants to use them.

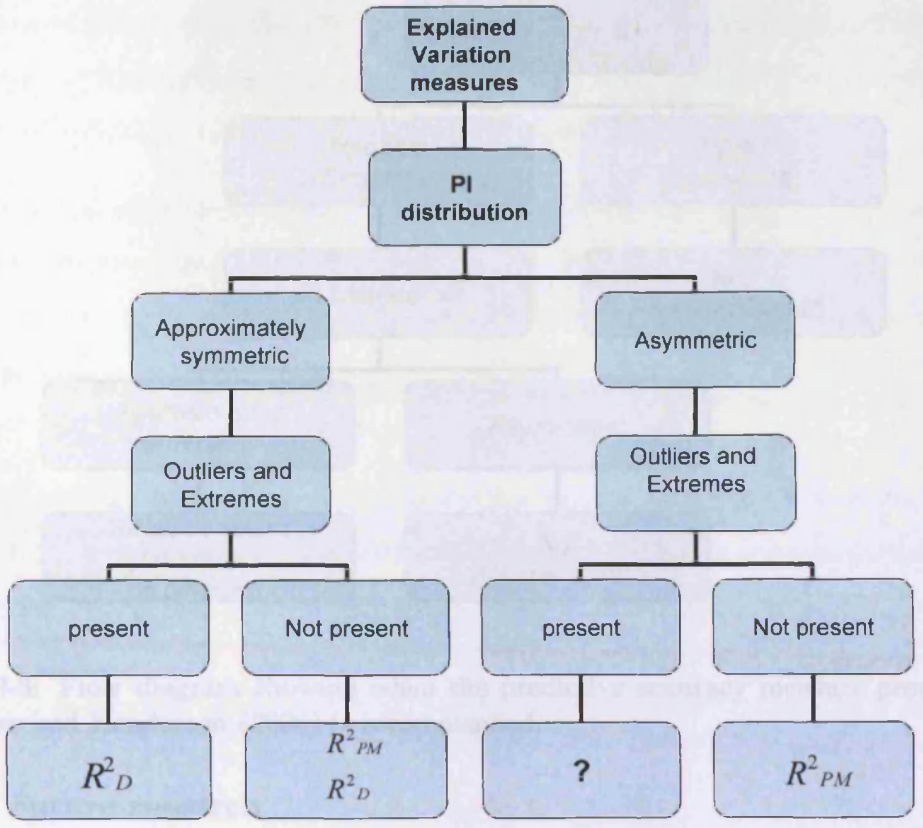


Figure 9-1: Flow diagram recommending an explained variation measure. Question mark: no measure is recommended.

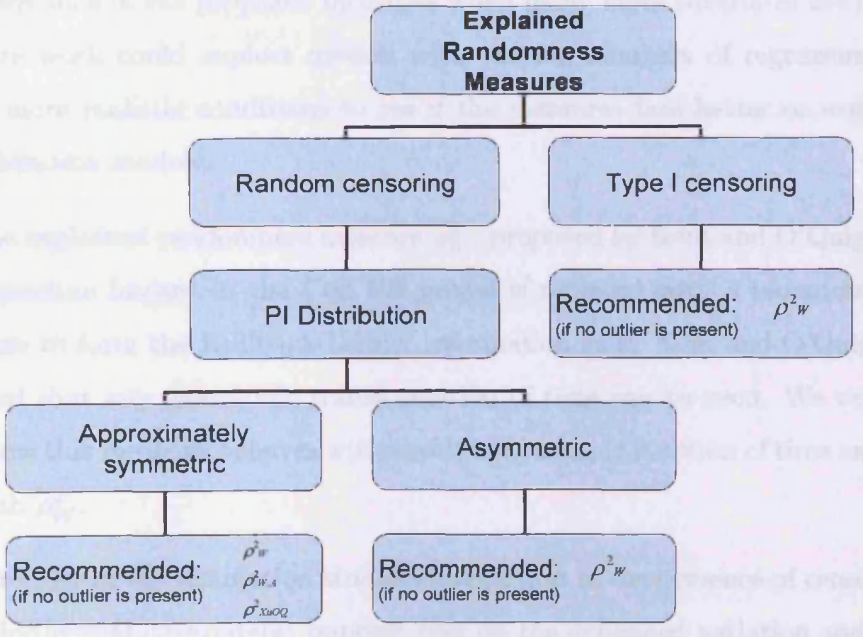


Figure 9-2: Flow diagram recommending an explained randomness measure.

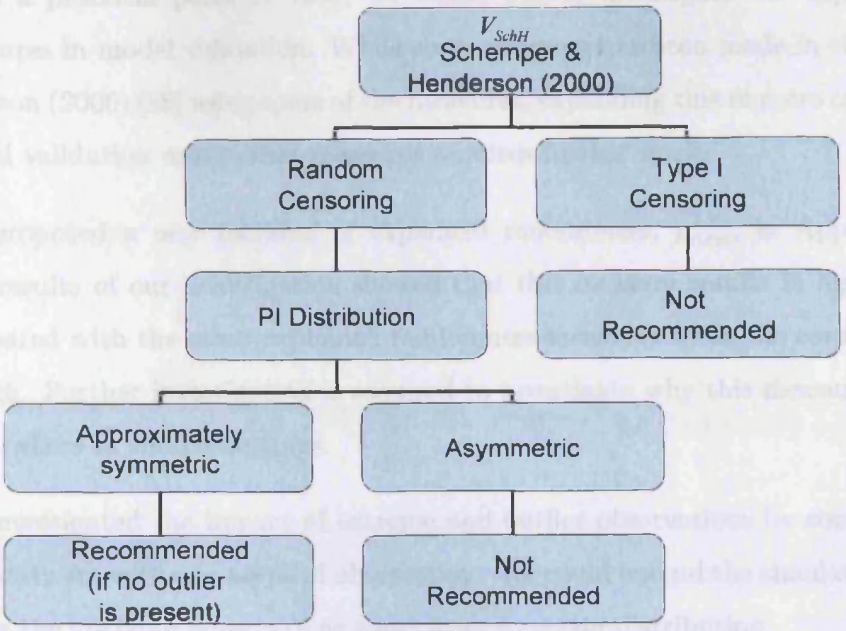


Figure 9-3: Flow diagram showing when the predictive accuracy measure proposed by Schemper and Henderson (2000) is recommended.

9.5.1 Future research

For further research we would like to expand on several studies:

1) Future simulations in this area could consider a wide variety of models to investigate performance of the proposed measures when many more covariates are introduced. Future work could explore models with varying numbers of regressors which reflect more realistic conditions to see if the measures fare better or worse in more cumbersome models.

2) In the explained randomness measure ρ^2_W, proposed by Kent and O'Quigley (1988), the baseline hazard in the Cox PH model is replaced with a monotonic function of time to form the Kullback-Leibler information gain. Kent and O'Quigley (1988) argued that any monotonic transformation of time can be used. We would like to see how this measure behaves with another monotonic function of time and compare it with ρ^2_W.

3) The results of the simulation studies showed that in the presence of censoring omitting influential covariate(s) imposes bias on the explained variation and explained randomness measures as well the estimated betas in the Cox PH model. We would like to investigate how to handle this difficult issue.

4) From a practical point of view, we would like to investigate the application of measures in model validation. While some progress has been made in this area by Royston (2006) [88] using some of the measures, expanding this to more complicated model validation using other measures requires further work.

5) We proposed a new measure of explained randomness, ρ^2_{new}, in Appendix B.8. The results of our investigation showed that this measure results in higher values compared with the other explained randomness measures when the covariate effect is high. Further investigation is required to investigate why this measure is higher than others in similar settings.

6) We investigated the impact of extreme and outlier observations by contaminating each data set with one atypical observation. We could extend the simulation studies where the outlying observations come from a certain distribution.

A number of areas have been identified in which the work in this thesis can be extended. These include the following:

7) Generalisability of the measures that have performed satisfactorily is one of the main areas that could be investigated. Royston and Sauerbrei (2004) [93] have proposed extensions to R^2_D for more flexible survival models. The performance of their proposed measures requires further investigation. Also, extending other measures requires further work.

8) The theoretical properties of the new explained randomness measure, ρ^2_{new}, in Appendix B.8 require further investigation.

9) Extension of the promising measures to partial and adjusted measure(s), similar to adjusted-R^2 in linear regression, is another area which would benefit from further investigation.

Appendix A

Simulation results by covariate distribution, censoring type, and censoring proportions

In this section, we present the results of simulation studies in more details. This section shows simulation results to study the impact of censoring on: I) explained variation measures; II) explained randomness measures; and III) predictive accuracy measures and R^2_{SchK}. The results are shown in similar tables to those of 5.6, 6.6, and 7.6. The tables indicate the performance of the measures in different covariate distributions, censoring mechanisms and censoring proportions. The figures in these tables are the average across four covariate effects, and three sample size conditions.

Table A.1: Summary performance of the explained variation measures proposed by Kent and O'Quigley (1988) and Royston and Sauerbrei (2004) by the covariate distribution, censoring mechanism, and censoring proportion.

Measure	Covariate Distribution		Random Censoring		Type I Censoring	
		% Censored	Average % Difference	*C.V.*	Average % Difference	*C.V.*
R^2_{PM}	normal	20	0.1	23.7	0.2	23.5
		50	0.5	28.6	0.5	28.1
		80	1.7	40.5	1.4	39.7
	lognormal	20	0.1	23.7	0.1	23.4
		50	0.4	27.5	0.3	26.8
		80	1.2	36.9	0.9	35.8
	pos. skewed	20	0.1	26.5	0.0	26.1
		50	0.3	28.9	0.1	28.1
		80	0.9	35.1	0.5	34.0
	neg. skewed	20	0.4	31.3	0.4	32.2
		50	1.1	38.7	1.5	39.9
		80	4.7	57.0	5.1	57.4
R^2_D	normal	20	0.1	23.8	0.2	23.6
		50	0.5	28.7	0.5	28.3
		80	1.9	40.7	1.5	40.0
	lognormal	20	4.2	24.7	5.5	24.4
		50	11.9	29.1	14.2	28.4
		80	24.0	39.6	26.3	38.6
	pos. skewed	20	12.8	31.5	15.9	30.5
		50	40.0	35.4	47.6	33.7
		80	88.5	43.8	97.7	41.9
	neg. skewed	20	-8.3	31.6	-13.3	31.7
		50	-19.7	39.2	-26.3	39.0
		80	-28.0	56.9	-31.5	55.9

Table A.2: Summary performance of the explained variation measures proposed by O'Quigley and Flandre (1994) and Xu and O'Quigley (2001) by the covariate distribution, censoring mechanism, and censoring proportion.

Measure	Covariate Distribution	% Censored	Random Censoring Average % Difference	Random Censoring *C.V.*	Type I Censoring Average % Difference	Type I Censoring *C.V.*
R^2_{OQF}	normal	20	2.2	23.3	2.6	22.9
		50	6.2	28.0	7.1	27.4
		80	13.0	39.5	13.4	38.7
	lognormal	20	3.1	23.0	2.9	22.8
		50	9.5	26.8	9.8	25.9
		80	21.3	36.5	21.5	35.2
	pos. skewed	20	2.7	34.4	0.7	32.2
		50	8.9	38.9	7.2	36.7
		80	24.1	39.0	22.0	60.1
	neg. skewed	20	-4.7	36.6	-9.6	31.4
		50	-10.7	41.1	-15.5	37.8
		80	-14.3	54.2	-16.6	52.8
R^2_{XuOQ}	normal	20	0.7	23.8	2.6	22.9
		50	3.3	36.1	7.1	27.4
		80	9.6	76.0	13.4	38.7
	lognormal	20	0.6	23.4	2.9	22.8
		50	3.3	29.9	9.9	25.9
		80	12.8	68.3	21.6	35.2
	pos. skewed	20	0.1	36.4	0.7	60.1
		50	0.3	54.3	7.2	32.2
		80	10.1	80.1	22.0	36.7
	neg. skewed	20	-1.1	43.1	-9.6	31.4
		50	-4.3	62.7	-15.5	37.8
		80	-9.2	90.1	-16.6	52.8

Table A.3: Summary performance of the explained variation measure proposed by Royston (2006) by the covariate distribution, censoring mechanism, and censoring proportion.

Measure	Covariate Distribution	% Censored	Random Censoring Average % Difference	Random Censoring *C.V.*	Type I Censoring Average % Difference	Type I Censoring *C.V.*
$R^2_{Royston}$	normal	20	5.5	24.0	7.9	23.8
		50	14.7	29.4	18.7	29.0
		80	28.6	42.4	33.1	41.9
	lognormal	20	9.7	23.7	13.2	23.5
		50	28.5	28.6	35.2	28.0
		80	60.2	40.0	68.5	39.3
	pos. skewed	20	15.2	25.0	19.3	24.4
		50	50.0	29.0	59.9	28.0
		80	117.3	38.1	131.0	36.8
	neg. skewed	20	-7.0	27.7	-10.9	27.9
		50	-18.4	34.6	-24.5	34.6
		80	-28.4	50.9	-31.0	50.3

Table A.4: Summary performance of the explained randomness measure proposed by Kent and O'Quigley (1988) by the covariate distribution, censoring mechanism, and censoring proportion.

Measure	Covariate Distribution	% Censored	Random Censoring Average % Difference	*C.V.*	Type I Censoring Average % Difference	*C.V.*
ρ^2_W	normal	20	0.1	22.1	0.1	21.9
		50	0.4	26.6	0.4	26.2
		80	1.4	37.6	1.1	36.8
	lognormal	20	0.1	21.4	0.1	21.0
		50	0.3	24.9	0.2	24.2
		80	1.0	33.6	0.7	32.6
	pos. skewed	20	0.1	22.5	0.0	22.1
		50	0.2	24.8	0.1	24.1
		80	0.7	30.8	0.3	29.7
	neg. skewed	20	0.4	35.5	0.3	36.2
		50	0.9	41.3	1.1	42.0
		80	3.2	54.8	3.3	54.7

Table A.5: Summary performance of the explained randomness measures proposed by Xu and O'Quigley (1999) and O'Quigley et al (2005) by the covariate distribution, censoring mechanism, and censoring proportion.

Measure	Covariate Distribution	% Censored	Random Censoring Average % Difference	*C.V.*	Type I Censoring Average % Difference	*C.V.*
ρ^2_{XuOQ}	normal	20	-0.1	22.4	6.3	22.2
		50	0.3	28.2	14.3	27.0
		80	4.8	44.0	24.3	38.6
	lognormal	20	0.1	22.0	10.8	21.9
		50	2.3	26.7	27.5	26.0
		80	15.4	41.6	50.5	36.0
	pos. skewed	20	0.2	23.5	16.0	23.2
		50	3.5	26.9	47.2	26.1
		80	30.3	40.8	95.8	33.7
	neg. skewed	20	-2.6	26.8	-10.0	26.3
		50	-9.1	35.5	-21.9	32.7
		80	-23.5	56.8	-27.8	47.4
ρ^2_k	normal	20	4.2	22.5	6.3	22.2
		50	11.1	27.4	14.3	27.0
		80	21.1	39.2	24.3	38.6
	lognormal	20	7.7	22.2	10.8	21.9
		50	22.1	26.6	27.5	26.0
		80	44.6	36.8	50.6	36.0
	pos. skewed	20	12.6	23.5	16.2	22.9
		50	39.7	27.0	47.6	25.9
		80	86.4	34.8	96.2	33.5
	neg. skewed	20	-6.2	26.1	-9.9	26.3
		50	-16.3	32.6	-21.8	32.7
		80	-25.4	47.9	-27.7	47.4

Table A.6: Summary performance of the predictive accuracy measures proposed by Graf et al (1999) and Schemper and Henderson (2000) by the covariate distribution, censoring mechanism, and censoring proportion. Note that the entries for the Graf's measure (1999) do not include 80% censoring.

			Random Censoring		Type I Censoring	
Measure	Covariate Distribution	% Censored	Average % Difference	*C.V.*	Average % Difference	*C.V.*
$R^2_G(T_4)$	normal	20	0.1	43.5	-2.2	44.4
		50	0.4	45.8	-7.0	46.7
		80				
	lognormal	20	0.2	41.4	-1.4	42.3
		50	0.5	43.6	-4.2	44.1
		80				
	pos. skewed	20	0.3	39.6	-0.4	40.4
		50	0.5	41.6	-1.0	41.8
		80				
	neg. skewed	20	-0.2	50.0	-7.2	51.9
		50	0.5	52.7	-23.6	55.2
		80				
V_{SchH}	normal	20	0.2	21.5	-1.0	21.1
		50	0.1	26.5	-13.8	25.4
		80	-6.9	42.8	-44.3	36.6
	lognormal	20	0.2	21.0	2.3	20.8
		50	1.5	25.0	-0.3	24.9
		80	2.2	37.6	-18.4	37.0
	pos. skewed	20	3.3	21.5	3.3	21.4
		50	12.0	26.4	11.9	26.3
		80	24.3	37.5	23.9	37.6
	neg. skewed	20	-6.2	25.4	-6.2	25.3
		50	-25.0	43.0	-24.6	43.0
		80	-50.0	83.2	-50.4	83.5

Table A.7: Summary performance of the measure proposed by Schemper and Kaider (1997) by the covariate distribution, censoring mechanism, and censoring proportion.

			Random Censoring		Type I Censoring	
Measure	Covariate Distribution	% Censored	Average % Difference	*C.V.*	Average % Difference	*C.V.*
R^2_{SchK}	normal	20	-0.3	25.8	0.1	25.2
		50	0.1	29.3	0.2	28.7
		80	1.7	39.7	1.7	39.1
	lognormal	20	-0.4	26.8	-0.1	26.2
		50	-0.1	29.2	-0.1	28.7
		80	1.3	36.7	1.2	35.8
	pos. skewed	20	-0.4	32.9	-0.1	32.6
		50	-0.2	33.4	0.0	33.5
		80	0.6	36.2	0.5	35.5
	neg. skewed	20	-0.5	33.4	-0.1	33.1
		50	0.1	37.9	0.5	38.3
		80	3.8	54.2	4.5	54.5

Appendix B

More details on some of the proposed measures

B.1 Royston and Sauerbrei D measure (2004)

Royston and Sauerbrei (2004) [93] define the D measure as follows. Suppose the data on n individuals are denoted by $(t_1, \delta_1, X_1), ..., (t_n, \delta_n, X_n)$ where for the ith individual t_i is the observed time, δ_i is 1 if the event of interest is experienced at t_i or 0 otherwise (right censoring), and X_i is the covariate vector of prognostic factors. The Cox model may be written as

$$\ln \lambda(t_i, X_i) = \ln \lambda_0(t_i) + h_i$$

where $h_i = \beta' X_i$ is the prognostic index (PI) for the ith individual. Consider the distribution of the PI values. Defining order statistics $h_{(1)} < ... < h_{(n)}$ we may quite generally write

$$h_{(i)} = \mu + \sigma u_i + \varepsilon_i \tag{B.1}$$

where u_i is the ith expected standard Normal order statistic (rankit) in a sample of size n. Ordering the data on the h_i and substituting for $h_{(i)}$ in B.1 we have (in an obvious notation)

$$\ln \lambda(t_{(i)}, X_{(i)}) = \ln \lambda_0(t_{(i)}) + \mu + \sigma u_i + \varepsilon_i$$

So far we have assumed no specific distribution for the h_i. Let us now suppose that the h_i are Normally distributed $N(\mu, \sigma^2)$. The parameter is the standard deviation ($S.D.$) of

the PI values and is a natural measure of separation. By definition, the regression of the $h_{(i)}$ on the u_i is linear with $E(h_{(i)}) = \mu + \sigma u_i$ and $E(\varepsilon_i) = 0$. To a first approximation, let us ignore the random perturbation i and set $i = 0$. Then

$$\ln \lambda(t_{(i)}, X_{(i)}) \simeq \ln \lambda_0(t_{(i)}) + \mu + \sigma u_i \tag{B.2}$$

Under the Normality assumption, the special Cox model B.2 is (approximately) linear in the u_i. On fitting it to the data, the constant is absorbed into the baseline hazard function and the regression coefficient, σ^* say, will estimate σ. Royston and Sauerbrei's proposed measure D is defined as

$$D = \kappa\sigma^* \tag{B.3}$$

where $\kappa = \sqrt{8/\pi} \simeq 1.60$.

B.1.1 Interpretation

D is log hazard ratio comparing two equal-sized prognostic groups based on dichotomising a continuous prognostic index ($\beta' X$).

B.2 $R(X)$ and R_0 in Korn and Simon measure (1990)

The measure proposed by Korn and Simon (1990) is

$$explained\ variation = \frac{R_0 - E_X[R(X)]}{R_0}.$$

The calculation of $R(X)$ and R_0 for loss function with squared error loss censored at T_0 is presented below. For this loss function there are essentially two predictions possible: survival less than T_0 or survival greater than or equal to T_0. One way to think about this loss function is that the time axis has been transformed so that the interval $[T_0, \infty)$ has been collapsed to the point T_0. From the definition of expected risk

$$R(X_i) = \int_0^\infty (t^* - \widehat{t}_{x_i})^2 dF(t|X_i)$$

and

$$R_0 = \int_0^\infty (t^* - \widehat{t})^2 dF(T)$$

where $t^* = \min(t, T_0)$, $\widehat{t}_{x_i} = E(T^* \mid X_i)$. To estimate these quantities we should replace $F(t|X_i)$ and the optimal predictor with their respective estimates, therefore

$$\begin{aligned} \widehat{R}(X_i) &= \int_0^\infty (\min(t, T_0) - E(T^*|X_i))^2 d\widehat{F}(t|X_i) \\ &= \int_0^{T_0} (t - E(T^*|X_i))^2 d\widehat{F}(t|X_i) + \int_{T_0}^\infty (T_0 - E(T^*|X_i))^2 d\widehat{F}(t|X_i) \end{aligned}$$

and $\widehat{t} = E(T^*)$ are the optimal predictors which are obtained as follows.

$$\begin{aligned} E(T^* \mid X_i) &= \int_0^\infty \min(t, T_0) dF(t|X_i) \\ &= \int_0^{T_0} t dF(t|X_i) + \int_{T_0}^\infty T_c dF(t|X_i) \\ &= \int_0^{T_0} t dF(t|X_i) + T_0(1 - F(T_0|X_i)) \end{aligned}$$

and similarly $E(T^*) = \int_0^{T_0} t dF(t) + T_0(1 - F(T_0))$ where $S(T_0|X_i) = 1 - F(T_0|X_i)$ and $S(T_0) = 1 - F(T_0)$. Expanding this and replacing $E(T^* \mid X_i) = \int_0^{T_0} t dF(t|X_i) + T_0(1 - F(T_0|X_i))$ will result in

$$\widehat{R}(X_i) = \int_0^{T_0} t^2 d\widehat{F}(t \mid X_i) + t^2 \widehat{S}(T_0 \mid X_i) - \left[\int_0^{T_0} t d\widehat{F}(t \mid X_i) + T_0 \widehat{S}(T_0 \mid X_i)\right]^2 \quad \text{(B.4)}$$

where

$$\int_0^{T_0} t^2 d\widehat{F}(t \mid X_i) = \sum_{t_j \leq T_0} t_j^2 \left[\widehat{S}(t_{j-} \mid X_i) - \widehat{S}(t_j \mid X_i)\right]$$

and

$$\int_0^{T_0} t d\widehat{F}(t \mid X_i) = \sum_{t_j \leq T_0} t_j \left[\widehat{S}(t_{j-} \mid X_i) - \widehat{S}(t_j \mid X_i)\right].$$

R_0 is also obtained in a similar manner by replacing $\widehat{F}(t \mid X_i)$ with $F_0(t) = \frac{1}{n}\sum_{i=1}^{n} F(t|x_i)$ and $\widehat{S}(T_0 \mid X_i)$ with $\widehat{S}(T_0) = \frac{1}{n}\sum_{i=1}^{n} S(T_0|x_i)$.

If a parametric model is used for survival data, then the explained risk will be a function of the unknown parameters. Substitution of consistent estimates of these parameter estimates will result in a consistent estimate of the explained risk. For example for the two-group exponential model with parameters λ_1 and λ_2 where $n_1 = n_2 = \frac{n}{2}$ depending on whether $x_i = 1$ or $x_i = 2$, the explained risk is: $\textit{explained risk} = \frac{\frac{1}{4\lambda_1^2} + \frac{1}{4\lambda_2^2} - \frac{1}{2\lambda_1\lambda_2}}{\frac{3}{4\lambda_1^2} + \frac{3}{4\lambda_2^2} - \frac{1}{2\lambda_1\lambda_2}}$.

B.3 Schemper and Kaider measure (1997)

Schemper and Kaider (1997) [98] proposed a measure, R^2_{SchK}, based on Spearman correlation coefficients and Kendall τ between survival times and covariates. They apply Rubin's multiple imputation method to augment censored survival times with random residual life times to make all survival times uncensored. Several such "augmented" data sets are generated and correlation between survival times (observed or completed/imputed) and independent variables are calculated using either Spearman correlation coefficients or Kendall τ for each data set then an average is taken. The algorithm is as follows.

1 We observe a sample data as (t_i, x_i), $i = 1, ..., n$, and estimate the parameters β and baseline survival function $S_0(t)$ for Cox's model with standardised covariate vector x. Note that $\widehat{S}_0(t)$ is only defined for $t < t^*$(t^* denotes the maximum observed uncensored life time).

2 Therefore we need to calculate an expected $\widehat{S}_0(t)$ for $t > t^*$, denoted by $\widehat{S}^e_0(t)$. To estimate $\widehat{S}^e_0(t)$, Schemper and Kaider (1997) [98] proposed a linear function which is fitted to the points $(t^*, \widehat{S}_0(t^*))$ and $(dt^*, 0)$ where d is a constant whose value can be chosen anything greater than 1, but for numerical reasons they recommend d = 2. therefore $\widehat{S}^e_0(t)$ will be

$$\widehat{S}^e_0(t) = b_0 + b_1 t$$

where $b_0 = \widehat{S}_0(t^*)\frac{d}{d-1}$ and $b_1 = -\widehat{S}_0(t^*)\frac{1}{t^*(d-1)}$. Due to assumed proportional hazards also the individual survival functions, $\widehat{S}_i(t) = \widehat{S}_0(t)^{\exp(\widehat{\beta}x_i)}$ (for $t < t^*$) and $\widehat{S}^e_i(t) = \widehat{S}^e_0(t)^{\exp(\widehat{\beta}x_i)}$ (for $t > t^*$) are now completely defined.

3 Next is to impute each censored survival time t^c_i become an uncensored time t_i according to the following procedure:

3.1 Draw a random number u_i, uniformly distributed in the interval $[0, \widehat{S}_i(t^c_i)]$ where $\widehat{S}_i(t^c_i) = \widehat{S}_0(t^c_i)^{\exp(\widehat{\beta}x_i)}$. Note that cumulative survival probabilities for $t > t^c_i$ are uniformly distributed in the interval $[0, \widehat{S}_i(t^c_i)]$ and that we draw u_i from one of these cumulative survival probabilities.

3.2 Then we follow the next steps to calculate t_i.

3.2.1 If $u_i \geqslant \widehat{S}_i(t^*)$ then $t_i = t_j$ for which $\widehat{S}_i(t_j) \geqslant u_i \geqslant \widehat{S}_i(t_{j+1})$ (t_j denotes the ordered uncensored survival times).

3.2.2 If $u_i < \widehat{S}_i(t^*)$ then $t_i = [\exp((\log u_i)/\exp(\widehat{\beta}x_i)) - b_0]/b_1$ with b_0 and b_1 calculated in step 2. The latter expression follows from equating $\widehat{S}_0^e(t_i)^{\exp(\widehat{\beta}x_i)} = (b_0 + b_1 t_i)^{\exp(\widehat{\beta}x_i)}$ to u_i.

4 Calculate a measure of correlation, R^2_{SchK}, using Spearman's correlation $r_s(T, X)$ or Kendall's $\tau(T, X)$ where T stands for either an observed or an imputed uncensored survival time.

5 Repeat steps 3 and 4 m times and then average R^2_{SchK}s to obtain $\overline{R}^2_{SchK}$. m is suggested to be $m \geqslant 3$.

B.4 Akazawa Measure (1997)

Akazawa (1997) [2] proposed a measure which is derived from the squared product-moment correlation; it can be interpreted as an adaptation of multiple correlation coefficient for normal linear model to the survival time regression model. He named his measure "MEVa". MEVa was proposed to calculate measure of explained variation in censored survival data with no loss to follow-up.

To explain his measure, let n be individuals entering a study at random over time and that the follow-up terminates at some prespecified time with no loss to follow-up. Let us consider the survival time setting. $T = \min(T_f, T_c)$ is the observed time, T_f is the survival time and T_c is the right censoring time. Let X_i be covariate vector for individual i. He defined three statistics e_i, $\overline{e}$ and $\overline{T}$ to use in his measure. He adapted equation (??) in simple linear regression to decomposed $T_i - \overline{T}$ into three components. In this method e_i, $\overline{e}$ and $\overline{T}$ are $e_i = E[T_i|X_i, T_c]$, $\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$ and $\overline{t} = \frac{1}{n}\sum_{i=1}^{n} T_i$ where $E[.|X_i, T_c]$ is the conditional expectation given X_i and T_c. Since $T_i - \overline{T} = (T_i - e_i) + (e_i - \overline{e}) + (\overline{e} - \overline{T})$, this follows

$$\frac{1}{n}\sum_{i=1}^{n}(T_i-\overline{T})^2 = \frac{1}{n}\sum_{i=1}^{n}(T_i-e_i)^2+\frac{1}{n}\sum_{i=1}^{n}(e_i-\overline{e})^2+\frac{2}{n}\sum_{i=1}^{n}(T_i-e_i)(e_i-\overline{e})+\frac{2}{n}(\overline{e}-\overline{T})\sum_{i=1}^{n}(T_i-\overline{e}) \tag{B.5}$$

Using the weak law of large numbers under suitable regularity conditions,

$\overline{T} - \overline{e} \longrightarrow 0$ and $\frac{1}{n}\sum_{i=1}^{n} T_i e_i - \frac{1}{n}\sum_{i=1}^{n} e_i^2 \longrightarrow 0$ as $n \longrightarrow +\infty$, in probability. Therefore, the expression (B.5) can be written asymptotically as $\frac{1}{n}\sum_{i=1}^{n}(T_i - \overline{T})^2 \longrightarrow \frac{1}{n}\sum_{i=1}^{n}(T_i - e_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(e_i - \overline{e})^2$ as $n \longrightarrow +\infty$, in probability.

Thus, the mean of total sum of squares about the mean of T_i is asymptotically decomposed into two parts: the means of the sum of squares about the mean and the sum of squares due to regression. Let t_i $(i = 1, ..., n)$ denote the observed time-on-study and $\overline{t}$ the mean of t_i. Then, the measure of explained variation is defined as

$$MEVa = \frac{\sum_{i=1}^{n}(e_i - \overline{e})^2}{\sum_{i=1}^{n}(t_i - \overline{t})^2} \tag{B.6}$$

The conditional expected time e_i lived in the time interval $[0, T_0]$ is estimated using the following expressions:

$$E(T) = \int_0^{Tc} T_f dF(t) + \int_{T_c}^{\infty} T_c dF(t)$$

and if we assume that $F(u|X_i, T_0)$ is the conditional distribution of $T_i = \min(T_f, T_0)$ defined on $[0, T_0)$ so that $F(u|X_i, T_0)$=$F(u|X_i)$ on $[0, T_0)$, and $F(u|X_i) = 1$ on $[T_0, \infty)$ then we can write:

$$\begin{aligned} e_i &= E[T|X_i, T_c] = \int_0^{T_c} u dF(u|X_i, T_c) = \int_0^{T_{c-}} u dF(u|X_i, T_c) + T_c\{1 - F(T_{c-}|X_i, T_c)\} \\ &= uF(u|X_i, T_c)|_0^{T_c} - \int_0^{T_{c-}} F(u|X_i, T_c)du + T_c\{1 - F(T_{c-}|X_i, T_c)\} \\ &= \int_0^{Tc} \{1 - F(u|X_i, T_c)\}du \end{aligned}$$

where T_{c-} is just the time before T_c. To calculate $MEVa$ in Cox proportional hazards regression model, the survival function $(1 - F(T_f|X_i, T_c))$ will be estimated by Link's methods (1984) [64], provided that the proportional hazards assumption holds.

B.5 Harrell measure (1986)

Harrell (1986) [37] defined a measure of explained variation for survival data. He introduced

$$R^2_{HL} = \frac{\log L(0) - \log L(b)}{\log L(0) - \log L^*} \tag{B.7}$$

as measure of explained variation for more general models where $\log L(0) - \log L(b)$ is likelihood ration test (LR) and L^* is the best (lowest) likelihood, so $\log L(0) - \log L^*$ is the amount of log-likelihood that is capable of being explained by the model. The lowest (best) possible log-likelihood for the Cox PH model is zero, so $\log L^*$ is zero in equation (B.7) for Cox model. He also introduced a measure similar to adjusted R^2 where the measure is penalised by the number of parameters in the model as

$$R^2_{adj-HL} = \frac{LR - 2p}{-2\log L(0)}. \tag{B.8}$$

The parameters in this measure are p, the number of parameters estimated and $LR = 2(\log L(b) - \log L(0))$. $L(b)$ and $L(0)$ are likelihoods of model with and without covariates. If the model LR is less than $2p$, R^2_{adj-HL} is set to 0.

B.6 Kent and O'Quigley measure (1988)

Heinzl (2000) [40] explored Kent and O'Quigley measure (1988), ρ^2_W, and introduced a SAS procedure to compute this measure. His proposed algorithm is presented below. The Cox proportional hazards model in (2.15) can be written as

$$f(t|X;\beta) = h_0(t)\exp\left\{\beta X - e^{\beta X}\int_0^t h_0(u)du\right\}. \tag{B.9}$$

Kent and O'Quigley (1988) [49] used $\rho^2_{IG} = 1 - \exp(-\Gamma(\beta_0))$ as a measure of explained variation where $\Gamma(\beta_0) = 2\{I(\beta_0;\beta_0) - I(0;\beta_0)\}$ and

$$I(\beta;\beta_0) = \int_X\int_T \log\{f(t|x;\beta)\}\, f(t|x;\beta_0)dtdG(x) \tag{B.10}$$

where $G(x)$ is the distribution function of X. In practice β_0 will be replaced by the maximum likelihood estimator $\widehat{\beta}$. Assuming no censoring a standard estimate of information gain will be provided by n^{-1} times the usual likelihood ratio test (Kent 1983 [50]). An alternative estimate, having similar statistical properties, is provided by the fitted information (Kent 1986 [51]) in which $I(\beta;\widehat{\beta})$, for $\beta = 0$ and $\beta = \widehat{\beta}$ are estimated by

$$I(\beta;\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_T \log\{f(t|x_i;\beta)\}\, f(t|x_i;\widehat{\beta})dt. \tag{B.11}$$

The distribution of X has been replaced by its empirical distribution in (B.11) . Kent and O'Quigley used (B.11) to form $I(\widehat{\beta};\widehat{\beta})$ and $I(0;\widehat{\beta})$ and then calculate ρ^2_{IG}. The detail of the procedure is explained below.

The problem with the Cox model is that the baseline hazard function $h_0(t)$ in (2.34) is completely unspecified. This makes it impossible to form $I(\beta;\widehat{\beta})$ in (B.11). To tackle this obstacle, Kent and O'Quigley used the following property of The Cox model. The inference and estimation of the parameters in the Cox model, as the result of using partial likelihood, is based on the survival time ranks not the actual survival times; therefore, any 'squeezing' or 'stretching' of the time axis does not change the results of the Cox regression model. It should neither change the result of a measure of dependence based on Cox regression model. Thus, any strictly monotone transformation of T, $T^* = \phi(T)$ gives the same Cox regression coefficient as T. Kent and O'Quigley utilized this property of the Cox model and defined $h_0^*(t) = \alpha\exp(\mu)t^{\alpha-1}$ for any choice of μ and α. By choosing this baseline hazard we can be ensured that the baseline hazard is proportional to power

of t. Therefore, if we replace $h_0(t)$ in (2.34) with $h_0^*(t)$, the conditional distribution T^* given $X = x$, $f^*(t|X;\beta)$, follows a Weibull distribution

$$f^*(t|X;\beta) = \alpha \exp(\mu + \beta X) t^{\alpha-1} \exp\left[-t^\alpha \exp(\mu + \beta X)\right],$$

and $Y = \ln T^*$ follows a linear regression model

$$Y = \ln(T^*) = -\sigma(\mu + \beta X) + \sigma\varepsilon \tag{B.12}$$

where $\sigma = \alpha^{-1}$ and ε is independent of X and has density $f(y)$ where

$$f(y) = e^y \exp(-e^y),$$

i.e. the extreme value (Gumbel) density (Lawless, 1982 [59]) with variance $\psi'(1) = 1.645$. σ and μ are scale and location parameters, respectively. Note that finding a suitable transformation would in practice not be possible if no parametric form for baseline hazard was assumed. Let $\theta = (\beta, \mu, \sigma^2)$ denote the parameters of the model. Let $\theta_1 = (\beta_1, \mu_1, \sigma_1^2)$ denote the true value of the parameters, generally with $\beta_1 \neq 0$. Define θ_1 to be the value of θ maximising the expected log likelihood, analogous to $\int_T \log\{f(t|x_i;\beta)\} f(t|x_i;\widehat{\beta})dt$ in (B.11), $\int_y \log(\{f(y|x;\theta)\} f(y|x;\theta_1)dy$ over θ satisfying H_0. Here $f(y|x;\theta_1) = \alpha f(\alpha y + \mu + \beta X)$ with $\alpha = \sigma^{-1}$.

Consider two hypothesis $H_0 : \theta_0 = (0, \mu_0, \sigma_0^2)$ and $H_1 : \theta_1 = (\beta, \mu_1, \sigma_1^2)$. It can be shown that ρ_{IG}^2 does not depend on the choice of μ_1, σ_1^2, they can be given arbitrary values. To make it as simple as possible, the best choice is μ_1 and $\sigma_1 = 1$ which corresponds to a constant baseline hazards function equal to one for H_1. The main problem is to find the estimator $\widehat{\theta}_0 = (0, \widehat{\mu}_0, \widehat{\sigma}_0^2)$. Just the appropriate values for μ_0 and $\sigma_0^2 > 0$ have to be computed.

It is obvious that the vector of the true model parameter values θ_1 is the θ maximising $\int_y \log(\{f(y|x;\theta)\} f(y|x;\theta_1)dy$ over all θ satisfying H_1. Therefore

$$\begin{aligned}\int_y \log(\{f(y|x;\theta)\} f(y|x;\theta_1)dy &= \int_{-\infty}^{+\infty} \alpha f(\alpha y + \mu + \beta X)\alpha_1 f(\alpha_1 y + \mu_1 + \beta_1 X)dy \\ &= \log(\alpha) + \frac{\alpha}{\alpha_1}\gamma'(1) + b - \exp(b)\gamma(\frac{\alpha}{\alpha_1} + 1)\end{aligned}$$

where $b = \mu + \beta x - (\frac{\alpha}{\alpha_1})(\mu_1 - \beta_1 x)$ and $\gamma(.)$ and $\gamma'(.)$ denote the gamma function and its derivative, respectively. This nonstandard notation is chosen to avoid confusion with

symbol Γ, which is already used for denoting information gain. The constant $\gamma'(1)$ =-0.577215...=$\psi(1)$ is the negative value of Euler's constant. Now we can estimate the measure of dependence between T and X , $\widehat{\rho}^2_W = 1 - \exp(-\widehat{\Gamma})$. Assume that for a study with n patients censored survival data have been observed with survival times t_i, censoring indicator c_i, and p-dimensional covariate vector x_i, $i = 1, .., n$. Fitting a Cox regression model under H_1 to the data, that is, using all p covariates yields estimated vector $\widehat{\beta}$ of regression coefficients. For calculating $\widehat{\rho}^2_W$ we consider $\theta_1 = (\widehat{\beta}, 0, 1)$ as true parameter values. Thus, $\widehat{\Gamma} = \Gamma\{ H_1 : H_0; \widehat{\theta}_1, G_n(x)\}$, where $G_n(x)$ denoted the empirical distribution of X putting mass $1/n$ at each of data points. To compute the estimator $\widehat{\theta}_0 = (0^T_p, \widehat{\mu}_0, \widehat{\alpha}_0)^T$ the empirical expected log likelihood $I(\theta; \widehat{\theta}_1) = \frac{1}{n} \sum_{i=1}^{n} \int_y \log(\{f(y|x;\theta)\} f(y|x;\theta_1) dy$ has to be numerically maximised with respect to μ and α, $\alpha > 0$. Taking partial derivatives and setting them to zero finally yields an explicit solution for $\widehat{\mu}_0$,

$$\widehat{\mu}_0 = -\log(\gamma(\widehat{\alpha}_0 + 1)) - \log(\frac{1}{n} \sum_{i=1}^{n} \exp(-\widehat{\alpha}_0 x_i \widetilde{\beta}))$$

and an implicit solution for $\widehat{\alpha}_0$,

$$\xi(\alpha) := \psi(1) - \psi(\alpha) + \sum_{i=1}^{n} \frac{\exp(-\alpha z_i)}{\sum_{j=1}^{n} \exp(-\alpha z_j)} z_j = 0$$

where $z_j = x_i \widehat{\beta} - \overline{x}\widehat{\beta}$, $i = 1, .., n$ and the vector $\overline{x}$ contain mean values of the p covariates. Heinzle (2000) showed how to solve numerical equation $\xi(\alpha)$ using (a) Newton-Raphson and (b) simple grid search. After we have found a numerical solution for $\widehat{\alpha}_0$, we can compute $\widehat{\Gamma} = 2\left\{ I(\theta; \widehat{\theta}_1) - I(\theta; \widehat{\theta}_0) \right\}$ and

$$\widehat{\rho}^2_W = 1 - \exp(-\widehat{\Gamma}) \tag{B.13}$$

where $\widehat{\Gamma} = 2[(1- \widehat{\alpha}_0)\psi(1) + \log\{\gamma(\widehat{\alpha}_0)\} + \log\{\frac{1}{n} \sum_{i=1}^{n} \exp(-\widehat{\alpha} z_i)\}]$.

B.7 Verweij and Van Houwelingen measure (1993)

Verweij and Houwelingen (1993) [113] proposed a similar measure to Magee's measure, $R^2_{LR} = 1 - \exp(-\frac{2}{n}(l(\beta) - l(0))$, in which the log-likelihood, $l(\beta)$, is replaced with the cross-validated log-likelihood, cvl.

The contribution of observation i to the log-likelihood can be defined as

$$l_i(\beta) = l(\beta) - l_{(-i)}(\beta),$$

where $l_{(-i)}(\beta)$ is the log-likelihood when observation i left out. The value of β that maximises $l_{(-i)}(\beta)$ is denoted by $\widehat{\beta}_{(-i)}$.

If the components of the likelihood are independent $l_i(\beta)$ simply equals the contribution of the ith component and

$$\sum_{i=1}^{n} l_i(\beta) = l(\beta).$$

Then the cross-validated log-likelihood cvl is defined by

$$cvl = \sum_{i=1}^{n} l_i(\widehat{\beta}_{(-i)}).$$

cvl can be considered as a measure of predictive value since for a given model cvl measures how well every observation i can be predicted using the other observations. For the computation of the cross-validated likelihood cvl, the coefficients $\widehat{\beta}_{(-i)}$ are required. They are estimated by fitting n models, each with $n - 1$ observations.

Verweij and Houwelingen (1993) [113] used cvl to define a cross-validated measure of explained variation in future data as

$$R^2_{cv} = 1 - \exp\left(-\frac{2}{n}(cvl - cvl_{null})\right).$$

B.8 A new measure of explained randomness for PH models

The measures of explained randomness ρ^2_W, ρ^2_{XuOQ}, and ρ^2_k make use of the properties of the Cox PH model [19] to quantify the predictive ability of the model. They are all based on the Kullback-Leibler information gain in equation 2.33, as discussed in section 2.3.2. Since the baseline hazard remains unspecified in the Cox PH model, the proposed explained randomness measures either replace the baseline hazard with a monotonic function of time, as in ρ^2_W, or work with the distribution of covariate(s) given time, as in ρ^2_{XuOQ} and ρ^2_k, to form the Kullback-Leibler information gain.

However, Ebrahimi and Kirmani (1996) [23] showed that Kullback-Leibler information gain ([55]) is independent of time for the proportional hazards models. We, therefore, develop the Kullback-Leibler information gain for the Cox PH model, and hence a new measure of explained randomness for the proportional hazards models.

The Cox PH model is defined as

$$\lambda(t|x) = \lambda(t).\exp(\beta' x)$$

and for simplicity in the maths operations consider $\alpha = \exp(\beta' x)$. Thus, the density function can be written as

$$f(t|x) = \lambda(t)\exp\{\beta' x - \alpha.\int_0^t \lambda(u)du\}$$

where

$$\Lambda(t) = \int_0^t \lambda(u)du \qquad S(t) = \exp\{-\int_0^t \lambda(u)du\} \qquad S(t|x) = S(t)^\alpha$$

$$f(t) = \lambda(t).S(t) \qquad f(t|x) = \lambda(t|x).S(t|x)$$

$$\lambda(t) = -\frac{S'(t)}{S(t)} \qquad \lambda(t|x) = -\frac{S'(t|x)}{S(t|x)}.$$

The Kullback-Leibler information gain is

$$KL = \int_0^\infty f(t|x) \ln(\frac{f(t|x)}{f(t)}) dt$$

and estimated information gain (Kent & O'Quigley (1988) [49]) for fitted density for T given X is

$$\textit{estimated} \quad KL = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty f(t|x_i) \ln(\frac{f(t|x_i)}{f(t)}) dt.$$

Therefore

$$\begin{aligned} KL &= \int_0^\infty f(t|x) \ln(\frac{f(t|x)}{f(t)}) dt \\ &= \int_0^\infty f(t|x) \ln(\frac{\lambda(t|x) S(t|x)}{\lambda(t) S(t)}) dt \end{aligned}$$

$$KL = \int_0^\infty f(t|x) \ln(\frac{\lambda(t|x)}{\lambda(t)}) dt + \int_0^\infty f(t|x) \ln(S(t|x)) dt - \int_0^\infty f(t|x) \ln(S(t)) dt$$

$$\int_0^\infty f(t|x) \ln(\frac{\lambda(t|x)}{\lambda(t)}) dt \qquad \text{(I)}$$

$$\int_0^\infty f(t|x) \ln(S(t|x)) dt \qquad \text{(II)}$$

$$\int_0^\infty f(t|x) \ln(S(t)) dt \qquad \text{(III)}$$

$$(I) = \int_0^\infty f(t|x) \ln(\frac{\lambda(t).\exp(\beta' x))}{\lambda(t)}) dt = \beta' x \int_0^\infty f(t|x) dt = \beta' x$$

$$
\begin{aligned}
(II)+(III) &= \int_0^\infty [f(t|x)\ln(S(t)^\alpha) - f(t|x)\ln(S(t))]\,dt \\
&= \int_0^\infty [f(t|x)(\alpha)\ln(S(t)) - f(t|x)\ln(S(t))]\,dt \\
&= \int_0^\infty (\alpha-1)(f(t|x).\ln(S(t))dt \\
&= (\alpha-1)\int_0^\infty f(t|x).\ln(S(t))dt \qquad \text{(IV)}
\end{aligned}
$$

$$
\begin{aligned}
\int_0^\infty f(t|x).\ln(S(t))dt &= \int_0^\infty \lambda(t|x).S(t|x).\ln(S(t))dt \\
&= \int_0^\infty -\frac{S'(t|x)}{S(t|x)}.S(t|x).\ln(S(t))dt \\
&= \int_0^\infty -S'(t|x).\ln(S(t))dt \\
&= -\int_0^\infty (S(t)^\alpha)'.\ln(S(t))dt \\
&= -\int_0^\infty \alpha S'(t)S(t)^{\alpha-1}.\ln(S(t))dt \\
S(t) &= z \qquad dz = S'(t)dt \qquad t:[0,\infty[\qquad z:[1,0[\\
&= \alpha\int_0^1 z^{\alpha-1}.\ln(z)dz
\end{aligned}
$$

$$
\int vdu = uv - \int udv \qquad v = \ln(z) \qquad dv = \frac{dz}{z} \qquad du = z^{\alpha-1}dz \qquad u = \frac{z^\alpha}{\alpha}
$$

$$
\begin{aligned}
&= \alpha.\left(\frac{1}{\alpha}z^\alpha.\ln(z)\Big|_0^1 - \int_0^1 \frac{1}{\alpha}z^{\alpha-1}dz\right) \\
&= \alpha.\left(\frac{1}{\alpha}z^\alpha.\ln(z)\Big|_0^1 - \frac{1}{\alpha.\alpha}z^\alpha\Big|_0^1\right) \\
&= \alpha.\left(\frac{1}{\alpha}1^\alpha.\ln(1) - \frac{1}{\alpha}0^\alpha.\ln(0)\right) - \alpha.\left(\frac{1}{\alpha.\alpha}1^\alpha - \frac{1}{\alpha.\alpha}0^\alpha\right)
\end{aligned}
$$

In the above equation, $\frac{1}{\alpha}0^\alpha.\ln(0)$ is of indeterminate form, so we should investigate

the limit

$$\lim_{y\to 0} y^{\alpha}.\ln(y).$$

We can write this limit as

$$\lim_{y\to 0} y^{\alpha}.\ln(y) = \lim_{y\to 0} \frac{\ln(y)}{\frac{1}{y^{\alpha}}}$$

which is now in the form ∞/∞, and hence the L'Hopital's Rule applies:

$$\lim_{y\to 0} \frac{\ln(y)}{\frac{1}{y^{\alpha}}} = \lim_{x\to 0} \frac{\frac{1}{y}}{\frac{1}{-\alpha.y^{\alpha+1}}} = \lim_{y\to 0} -\alpha.y^{\alpha} = 0$$

hence

$$\alpha.\left(\frac{1}{\alpha}.1^{\alpha}.0 - \frac{1}{\alpha}.0\right) - \alpha.\left(\frac{1}{\alpha.\alpha} - 0\right) =$$

$$= -\alpha.\left(\frac{1}{\alpha.\alpha}\right) = -\frac{1}{\alpha}$$

$$(IV) = (\alpha - 1).\frac{-1}{\alpha} = \frac{1}{\alpha} - 1$$

thus

$$KL = \beta' x - 1 + \frac{1}{\alpha} = \beta' x - 1 + \exp(-\beta' x)$$

and the estimated information gain (Kent & O'Quigley (1988) [49]) is

$$estimated \quad KL = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\beta}' x_i - 1 + \exp(-\widehat{\beta}' x_i)\right] \tag{B.14}$$

where x_i, $i = 1, 2, ...n$, is the covariate, and $\widehat{\beta}$ is the maximum likelihood estimator of the parameter in the model. In the case of multiple regression, $\widehat{\beta}' x_i$ will be replaced with prognostic index, i.e. linear predictor, of the model to compute estimated information gain.

Finally, using the relationship between the Kullback-Leibler information gain and the measures of predictive ability proposed by Kent & O'Quigley (1988) [49], presented in equation (2.33), a new measure for the non-stratified proportional hazards model is defined as

$$New\ measure \quad \rho^2_{new} = 1 - \exp\left\{-2\left(\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\beta}' x_i - 1 + \exp(-\widehat{\beta}' x_i)\right]\right)\right\}.$$

If the covariate, X, is normally distributed, $N(0,1)$, the expected value of KL with respect to X is

$$E_x\left(KL\right) = \beta E(X) - 1 + E_x\left[\exp(-\beta x)\right]$$

where

$$\begin{aligned} E_x\left[\exp(-\beta X)\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\beta X} e^{-X^2/2} dX \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-X^2/2-\beta X} dX \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\left[-\frac{(X+\beta)^2}{2}+\frac{\beta^2}{2}\right]} dX \\ &= e^{\frac{\beta^2}{2}}.\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(X+\beta)^2}{2}} dX \\ &= e^{\frac{\beta^2}{2}} \end{aligned}$$

since $E(X) = 0$, therefore

$$E_{X\sim N(0,1)}(KL) = -1 + e^{\frac{\beta^2}{2}}.$$

Then a measure of explained randomness for the univariate model where $X \sim N(0,1)$ is

$$New\ \rho^2_{New|X\sim N(0,1)} = 1 - \exp\left\{-2\left(-1 + e^{\frac{\beta^2}{2}}\right)\right\}.$$

The new measure can be evaluated for different β values in this univariate model as follows

Hazard Ratio	β	*New* $\rho^2_{New\|X\sim N(0,1)}$	ρ^2_W	$\rho^2_{W,A}$	ρ^2_{XuOQ}	ρ^2_k
1.25	0.223	0.049	0.049	0.050	0.048	0.048
1.5	0.405	0.157	0.141	0.143	0.134	0.134
2	0.693	0.419	0.316	0.325	0.296	0.296
4	1.386	0.960	0.637	0.657	0.602	0.602

B.8.1 Extension to the stratified Cox PH model

Kullback-Leibler information gain in equation 2.33 can be modified for the stratified Cox PH model. In the stratified model, the Kullback-Leibler information gain in equation 2.33 can be measured by twice the weighted average of the stratum-specific information gains

$$KL = \sum_{s=1}^{m} \frac{n_s}{n} \int_0^{\infty} f(t|x,s) \ln(\frac{f(t|x,s)}{f(t)}) dt.$$

By repeating the same maths operations, similar equation to B.14 can be derived for the stratified Cox models where, for example, the variable X_2 is split into strata which are represented by the m-level-factor S

$$\begin{aligned} KL_s &= \sum_{s=1}^{m} \frac{n_s}{n} \left\{ \frac{1}{n_s} \sum_{i=1}^{n_s} \left[\beta' x_{si} - 1 + \exp(-\beta' x_{si}) \right] \right\} \\ &= \frac{1}{n} \sum_{s=1}^{m} \sum_{i=1}^{n_s} \left[\beta' x_{si} - 1 + \exp(-\beta' x_{si}) \right] \end{aligned}$$

where $s = 1, 2, ..., m$ and n_s is the number of observations in strata s. For the stratified Cox PH model, KL_s replaces KL in ρ^2_{new} to provide a new measure of explained randomness for the stratified Cox PH model.

In summary, a measure of explained randomness for the Cox PH model can be defined without replacing the baseline hazard with a monotonic transformation of time, as in ρ^2_W, or reversing the role of outcome and covariate, as in ρ^2_{XuOQ}. In normally distributed covariates, the new measure ρ^2_{New} is in agreement with other explained randomness measures in small to moderate covariate effects, but it results in much higher values if the covariate effect is large, i.e. 1.386. This new measure is independent of censoring, intuitive, and can be modified for the stratified Cox PH model.

Appendix C

Models fitted to data sets in chapter 8

C.1 Models fitted to leg ulcer study data set

C.1.1 MFP I model:

```
. fracgen ulcarea 0.5
. fracgen mthson 0, replace
. fracgen age -2

. stcox age_1 mthson_1 ulcare_1  diastbp deepppg, nohr

No. of subjects =          200         Number of obs    =       200
No. of failures =           97
Time at risk    =        14232
                                       LR chi2(5)       =    119.89
Log likelihood  =   -387.77022         Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_1   26.04694   7.32928     3.55   0.000     11.68182    40.41207
    mthson_1    -.45057    .1030608  -4.37   0.000     -.65257    -.2485791
```

```
ulcare_1   -8.49743  1.46874    -5.79   0.000 -11.37612  -5.618759
 diastbp    -.01884   .0081183  -2.32   0.020   -.03476   -.0029353
 deepppg    -.58603   .2096234  -2.80   0.005   -.99689   -.1751824
------------------------------------------------------------------------------
```

C.1.2 MFP I model after removing 5 extreme observations:

```
. stcox age_1 mthson_1 ulcare_1  diastbp deepppg, nohr

No. of subjects =          195         Number of obs    =       195
No. of failures =           97
Time at risk    =        13620
                                       LR chi2(5)       =    109.81
Log likelihood  =    -387.7683         Prob > chi2      =    0.0000

------------------------------------------------------------------------------
  _t       Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
 age_1   26.04648  7.329038     3.55   0.000    11.68183  40.41113
mthson_1  -.45053   .103060    -4.37   0.000    -.65252   -.24853
ulcare_1 -8.49398  1.470568    -5.78   0.000  -11.37625  -5.61172
diastbp   -.01884   .008118    -2.32   0.020    -.03475   -.00293
deepppg   -.58604   .209622    -2.80   0.005    -.99689   -.17519
------------------------------------------------------------------------------
```

C.1.3 MFP II model:

```
. stcox age mthson_1 ulcarea  diastbp deepppg, nohr

No. of subjects =          200         Number of obs    =       200
No. of failures =           97
Time at risk    =        14232
                                       LR chi2(5)       =    113.74
Log likelihood  =   -390.84948         Prob > chi2      =    0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0212034	.0087131	-2.43	0.015	-.038	-.0041261
mthson_1	-.4391165	.1029093	-4.27	0.000	-.640	-.237418
ulcarea	-.0016209	.0003338	-4.86	0.000	-.002275	-.00097
diastbp	-.0178674	.0080573	-2.22	0.027	-.0336594	-.00208
deepppg	-.5714134	.2100124	-2.72	0.007	-.9830302	-.15980

C.1.4 MFP II model after removing 5 extreme observations:

```
. stcox age mthson_1 ulcarea  diastbp deepppg, nohr
```

No. of subjects =	195	Number of obs =	195
No. of failures =	97		
Time at risk =	13620		
		LR chi2(5) =	103.64
Log likelihood =	-390.84948	Prob > chi2 =	0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0212034	.0087131	-2.43	0.015	-.0382808	-.0041261
mthson_1	-.4391165	.1029093	-4.27	0.000	-.640815	-.237418
ulcarea	-.0016209	.0003338	-4.86	0.000	-.002275	-.0009667
diastbp	-.0178674	.0080573	-2.22	0.027	-.0336594	-.0020753
deepppg	-.5714134	.2100124	-2.72	0.007	-.9830302	-.1597966

C.2 Models fitted to breast cancer I study data set

```
. stset rfs rfsstat
```

C.2.1 RFS I model:

```
. stcox age er gradd1 gradd2 size nodd1 nodd2 ther1, nohr

No. of subjects =          295        Number of obs    =       295
No. of failures =          118
Time at risk    =  24975.06033
                                      LR chi2(8)       =     50.51
Log likelihood  =   -597.04107        Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0522733   .0162416    -3.22   0.001    -.0841063   -.0204404
          er |  -.4389279    .219263    -2.00   0.045    -.8686755   -.0091802
      gradd1 |   .8948001   .3107595     2.88   0.004     .2857227    1.503878
      gradd2 |   .9299319   .3194396     2.91   0.004     .3038419    1.556022
        size |   .3372146   .1943897     1.73   0.083    -.0437823    .7182114
       nodd1 |   .2771259   .3082003     0.90   0.369    -.3269356    .8811873
       nodd2 |   .8061004   .3546643     2.27   0.023     .1109712     1.50123
       ther1 |  -.5869892   .3056888    -1.92   0.055    -1.186128    .0121498
------------------------------------------------------------------------------
```

C.2.2 RFS II model:

```
. stcox age er gradd1 gradd2 size nodd1 nodd2 ther1 gene70, nohr

No. of subjects =          295        Number of obs    =       295
No. of failures =          118
Time at risk    =  24975.06033
                                      LR chi2(9)       =     72.62
Log likelihood  =   -585.98855        Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------+------------------------------------------------------------
   age |  -.0447666   .0162788    -2.75   0.006    -.0766725   -.0128607
    er |  -.1565484   .2169731    -0.72   0.471    -.5818079    .2687111
gradd1 |   .4479252   .3270106     1.37   0.171    -.1930039    1.088854
gradd2 |   .2809455   .3421952     0.82   0.412    -.3897448    .9516358
  size |   .3665994   .1943658     1.89   0.059    -.0143505    .7475493
 nodd1 |   .1836802   .3037698     0.60   0.545    -.4116977    .7790581
 nodd2 |    .781884   .3649559     2.14   0.032     .0665835    1.497184
 ther1 |  -.6095437   .3082959    -1.98   0.048    -1.213793   -.0052948
gene70 |   1.236802   .2822552     4.38   0.000     .6835923    1.790012
------------------------------------------------------------------------
```

C.2.3 OS I model:

```
. stset os osstat

. stcox age er gradd1 gradd2 size nodd1 nodd2 ther1, nohr

No. of subjects =          295                 Number of obs    =       295
No. of failures =           79
Time at risk    =  27838.31012
                                                LR chi2(8)       =     60.61
Log likelihood  =   -387.99393                  Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0402315    .019744    -2.04   0.042     -.078929    -.001534
          er |  -.8276249   .2504392    -3.30   0.001    -1.318477   -.3367731
      gradd1 |   1.460347     .54231     2.69   0.007     .3974392    2.523255
      gradd2 |   1.785506   .5409229     3.30   0.001     .7253168    2.845696
        size |   .4154089   .2417264     1.72   0.086     -.058366    .8891839
       nodd1 |   .0624291   .4063887     0.15   0.878     -.734078    .8589362
       nodd2 |   .6153807   .4397907     1.40   0.162    -.2465932    1.477355
       ther1 |  -.2107752   .3912291    -0.54   0.590    -.9775701    .5560197
```

C.2.4 OS II model:

```
. stcox age er gradd1 gradd2 size nodd1 nodd2 ther1 gene70, nohr

No. of subjects =          295          Number of obs    =       295
No. of failures =           79
Time at risk    =  27838.31012
                                         LR chi2(9)       =     77.64
Log likelihood  =   -379.47708           Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0338659   .0196702    -1.72   0.085    -.0724187     .004687
          er |  -.5339754   .2475778    -2.16   0.031    -1.019219   -.0487317
      gradd1 |   .9378675   .5580181     1.68   0.093    -.1558278    2.031563
      gradd2 |    1.04331   .5623386     1.86   0.064    -.0588536    2.145473
        size |   .4569345   .2413038     1.89   0.058    -.0160123    .9298812
       nodd1 |   .0125814   .3975092     0.03   0.975    -.7665224    .7916852
       nodd2 |   .6759106   .4547843     1.49   0.137    -.2154502    1.567271
       ther1 |  -.2929577   .3955662    -0.74   0.459    -1.068253    .4823379
      gene70 |   1.550349   .4322426     3.59   0.000     .7031692    2.397529
------------------------------------------------------------------------------
```

C.3 Models fitted to breast cancer II study data set

```
. describe hormon x1 x2 x3 x4 x5 x6 x7 rectime censrec

obs:           686                          German breast cancer dataset
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
```

```
-------------------------------------------------------------------------------
hormon          byte    %12.0g      Therapy     Hormonal Therapy
x1              byte    %9.0g                   Age
x2              byte    %14.0g      menop       Menopausal status
x3              int     %9.0g                   Tumour size
x4              byte    %9.0g                   Tumour grade
x5              byte    %9.0g                   Number of positive nodes
x6              int     %9.0g                   Progesterone receptor
x7              int     %9.0g                   Estrogen receptor
rectime         int     %9.0g                   Recurrence free survival time
censrec         byte    %9.0g       cencode     Censoring Indicator
-------------------------------------------------------------------------------

. stset  rectime censrec

. fracgen x1 -2 -.5
. fracgen x6 .5
. gen x5a=cond(x5>=3,1,0)
. gen x5b=cond(x5>=9,1,0)
. gen x4a=cond(x4>=2,1,0)
. gen x5e=exp(-0.12*x5)
```

C.3.1 Linear model:

```
. stcox x4a x5a x5b x6_1 hormon, nohr

No. of subjects =          686          Number of obs    =       686
No. of failures =          299
Time at risk    =       771400
                                        LR chi2(5)       =    122.90
Log likelihood  =   -1726.7229          Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
   x4a |      .60101      .24902     2.41   0.016      .1129398     1.08908
   x5a |    .5943838    .1376574     4.32   0.000      .3245803    .8641874
   x5b |     .601952    .1441665     4.18   0.000      .3193909    .8845131
   x6a |   -.0570375    .0111694    -5.11   0.000     -.0789292   -.0351459
hormon |   -.3842406    .1252575    -3.07   0.002     -.6297408   -.1387405
------------------------------------------------------------------------------
```

C.3.2 MFP model:

```
. stcox x1_1 x1_2 x4a x5e x6_1 hormon, nohr

No. of subjects =          686         Number of obs    =       686
No. of failures =          299
Time at risk    =       771400
                                       LR chi2(6)       =    153.11
Log likelihood  =   -1711.6186         Prob > chi2      =    0.0000

------------------------------------------------------------------------------
    _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
   x1a |    1.742153   .3301373     5.28   0.000      1.095095     2.38921
   x1b |   -7.817902   1.749447    -4.47   0.000     -11.24675   -4.389049
   x4a |    .5174351   .2493739     2.07   0.038      .0286713    1.006199
   x5e |   -1.981213   .2268903    -8.73   0.000     -2.425909   -1.536516
   x6a |   -.0581884   .0110946    -5.24   0.000     -.0799335   -.0364433
hormon |   -.3944998    .128097    -3.08   0.002     -.6455653   -.1434342
------------------------------------------------------------------------------
```

C.4 Model fitted to prostate cancer study data set

```
obs:           506                          Bone prostate data

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
```

```
ap              int    %8.0g                 Acid phosphatase
pf              byte   %8.0g                 Performance status
------------------------------------------------------------------------------
Sorted by:

. fracgen ap 0

. stcox ap_1 pf age wt hx hg sz, nohr

No. of subjects =          506              Number of obs    =       506
No. of failures =          356
Time at risk    =        18551
                                            LR chi2(7)       =     77.41
Log likelihood  =    -1990.015              Prob > chi2      =    0.0000

------------------------------------------------------------------------------
    _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
  ap_1 |   .0642943   .0325554     1.97   0.048      .000487    .1281017
    pf |   .3785905   .1595785     2.37   0.018     .0658224    .6913585
   age |   .0211054   .0083726     2.52   0.012     .0046954    .0375154
    wt |  -.0108855   .0043808    -2.48   0.013    -.0194718   -.0022993
    hx |   .4718573    .110972     4.25   0.000     .2543562    .6893584
    hg |  -.0068701   .0029825    -2.30   0.021    -.0127156   -.0010245
    sz |   .0164314   .0043956     3.74   0.000     .0078162    .0250467
------------------------------------------------------------------------------
```

C.5 Models fitted to renal cancer I study data set

```
describe  t_mt who2 who3  haem iwcc1 trt

              storage  display     value
variable name   type   format      label  variable label
------------------------------------------------------------------------------
```

```
t_mt            int    %8.0g                 Days from diagnosis of
                                             metastasis to randomisation
who2            byte   %8.0g                 WHO PS 1
who3            byte   %8.0g                 WHO PS 2
haem            float  %9.0g                 HAEMOGLOBIN
iwcc1           float  %9.0g
trt             int    %8.0g       trt_      TREATMENT
```

C.5.1 Linear model:

```
. stcox it_mt1 who2 who3 ihaem1 iwcc1 trt, nohr

No. of subjects =          347          Number of obs    =       347
No. of failures =          322
Time at risk    =  4507.752957
                                        LR chi2(6)       =    122.71
Log likelihood  =   -1552.1855          Prob > chi2      =    0.0000

------------------------------------------------------------------------------
         _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
      t_mt1 |  -.0003498   .0002028    -1.73   0.084    -.0007473    .0000476
       who2 |   .2527672   .1396847     1.81   0.070    -.0210097    .5265441
       who3 |   .833791    .1657898     5.03   0.000     .5088489   1.158733
      haem1 |  -.2152997   .0345636    -6.23   0.000    -.2830431   -.1475563
       wcc1 |   .0696692   .0132308     5.27   0.000     .0437374    .095601
       trt  |  -.3491271   .113224     -3.08   0.002    -.571042    -.1272122
------------------------------------------------------------------------------
```

C.5.2 MFP model:

```
. fracgen it_mt1 -.5

. stcox it_mt1_1 who2 who3 ihaem1 iwcc1 trt, nohr
```

```
No. of subjects =          347          Number of obs    =       347
No. of failures =          322
Time at risk    =  4507.752957
                                        LR chi2(6)       =    132.69
Log likelihood  =   -1547.1966          Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    t_mt1_1 |   .0441056   .0108671     4.06   0.000     .0228065    .0654047
      who2 |   .2855838   .1378819     2.07   0.038     .0153402    .5558274
      who3 |   .8871175   .1639929     5.41   0.000     .5656974    1.208538
     haem1 |  -.2067895   .0344722    -6.00   0.000    -.2743538   -.1392252
      wcc1 |     .06934   .0132359     5.24   0.000     .0433981     .095282
       trt |  -.3338317   .1130017    -2.95   0.003     -.555311   -.1123525
------------------------------------------------------------------------------
```

C.6 Models fitted to renal cancer II study data set

```
. fracgen crp -2
```

```
. des  age_t lk_t liver_t bone_t neutr_ul  crp_t
```

```
              storage  display     value
var name       type    format      label      variable label
------------------------------------------------------------------------------
age            byte    %8.0g
lk             byte    %8.0g       lk         lymph node metastasis
liver          byte    %8.0g       liver      liver metastasis
bone           byte    %8.0g       bone       bone metastasis
neutr_ul       int     %8.0g                  neutrophils
crp            int     %8.0g                  crp-protein
```

```
. stcox  age lk liver bone neutr_ul  crp, nohr

No. of subjects =          322            Number of obs    =       322
No. of failures =          274
Time at risk    =  10399.67147
                                          LR chi2(6)       =     48.78
Log likelihood  =   -1365.6732            Prob > chi2      =    0.0000

------------------------------------------------------------------------------
    _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
   age |  -.0206172   .0072866    -2.83   0.005    -.0348986   -.0063358
    lk |   .3321942   .1306681     2.54   0.011     .0760894     .588299
 liver |   .3162244   .1725015     1.83   0.067    -.0218724    .6543211
  bone |   .6104281   .1528271     3.99   0.000     .3108925    .9099636
neutr_ul   .0001386   .0000361     3.84   0.000     .0000678    .0002094
crp_1  |  -8.035771   3.726544    -2.16   0.031    -15.33966   -.7318786
------------------------------------------------------------------------------
```

C.7 Model fitted to PBC I study data set

```
. gen cens=cond(status==2,1,0)
. gen age_y=age/365.25
. gen lnalbumin=log(albumin)
. gen lnbilir=log(bilir)
. gen lnpro_time=log(pro_time)

. stset time cens

. stcox age_y edema lnalbumin lnbilir lnpro_time, nohr

No. of subjects =          312            Number of obs    =       312
No. of failures =          125
Time at risk    =       625985
```

```
                                        LR chi2(5)      =     199.13
Log likelihood  =   -540.41244          Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_y |    .033266   .0086598     3.84   0.000      .0162933    .0502391
       edema |    .784686   .2991328     2.62   0.009      .1983967    1.370976
   lnalbumi |  -3.053267   .7240783    -4.22   0.000     -4.472434   -1.634099
     lnbilir |    .879207   .0987322     8.90   0.000      .6856962    1.072719
  lnpro_tim |   3.015681   1.023797     2.95   0.003      1.009076    5.022286
------------------------------------------------------------------------------
```

C.8 Model fitted to PBC II study data set

```
 . stset  time dead

. describe age  bilir cirrh central treat

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
age             float  %9.0g                  Age
bilir           float  %9.0g                  Bilirubin
cirrh           int    %8.0g                  Cirrhosis [1=yes]
central         int    %8.0g                  Central cholestasis [1=yes]
treat           int    %8.0g                  Treatment

. fracgen  bilir 0

. stcox age  bilir_1 cirrh central treat, nohr

No. of subjects =          207          Number of obs    =       207
No. of failures =          105
```

```
Time at risk     =       313913
                                            LR chi2(5)      =     136.81
Log likelihood   =   -422.14087             Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .046440   .0109895     4.23   0.000      .024901     .067979
     bilir_1 |   1.078658   .1300643     8.29   0.000      .823736    1.333579
      cirrh |    .924442   .214637      4.31   0.000      .503761    1.345123
     central |    .769449   .2657023     2.90   0.004      .248683    1.290217
       treat |   -.498985   .2016777    -2.47   0.013     -.894266    -.103704
------------------------------------------------------------------------------
```

C.9 Model fitted to lymphoma study data set

C.9.1 Model I:

```
. stcox ipi_dd1 ipi_dd2, nohr

No. of subjects =           73       Number of obs    =        73
No. of failures =           48
Time at risk    =  332.0099985
                                     LR chi2(2)       =      7.55
Log likelihood  =    -176.8355       Prob > chi2      =    0.0229

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ipi_dd1 |   .9873513    .458733     2.15   0.031      .0882512    1.88645
     ipi_dd2 |   .8289193    .348052     2.38   0.017      .1467487    1.51109
------------------------------------------------------------------------------
```

C.9.2 Model II:

```
. stcox ipi_dd1 ipi_dd2 outcome_predictor_score, nohr

No. of subjects =          73         Number of obs    =        73
No. of failures =          48
Time at risk    =  332.0099985
                                      LR chi2(3)       =     17.64
Log likelihood  =   -171.79222        Prob > chi2      =    0.0005

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ipi_dd1 |   1.042206   .4620482     2.26   0.024      .136608    1.947804
     ipi_dd2 |    .718093   .3509191     2.05   0.041      .030305    1.405883
     gene_17 |    .719341   .2268566     3.17   0.002      .274710    1.163972
------------------------------------------------------------------------------
```

Bibliography

[1] R. P. Abelson. A variance explanation paradox - when a little is a lot. *Psychological Bulletin*, 97(1):129–133, 1985.

[2] K. Akazawa. Measures of explained variation for a regression model used in survival analysis. *Journal of Medical Systems*, 21(4):229–238, 1997.

[3] P. D. Allison. *Survival Analysis Using the SAS System.* SAS Institute Inc., Carolina, U.S.A., 1995.

[4] D. G. Altman and P. Royston. Criteria for the validation of surrogate endpoints in randomized experiments. *Statistics in Medicine*, 19:453–473, 2000.

[5] D. G. Altman and P. Royston. The cost of dichotomising continuous variables. *BMJ*, 332:1080, 2006.

[6] R. Anderson-Sprecher. Model comparisons and R2. *The American Statitician*, 48(2):113–117, 1994.

[7] D. F. Andrew and A. M. Herzberg. *Data.* Springer, Berlin, 1985.

[8] J. Atzpodien, P. Royston, and M. Reitz. Metastatic renal carcinoma extended staging system. *British Journal of Cancer*, 88(3):348–353, 2003.

[9] V. Barnett and T. Lewis. *Outliers in Statistical Data.* Wiley and Sons, New York, 3 edition, 1994.

[10] F. M.-S. Barthel, A. Babiker, P. Royston, and M. K. B. Parmar. Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. *Statistics in Medicine*, 25(15):2521–2542, 2006.

[11] R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

[12] A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25:4279–4292, 2006.

[13] M. Buyse and G. Molenberghs. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, 54:1014–1029, 1998.

[14] D. P. Byar and S. B. Green. The choice of treatment for cancer patients based on covariate information - application to prostate cancer. *Bulletin Du Cancer*, 67(4):477–490, 1980.

[15] E. Christensen, J. Neuberger, J. Crowe, D.G Altman, H. Popper, B. Portmann, D. Doniach, L. Ranek, N. Tygstrup, and R. Williams. Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: final results of an international trial. *Gastroenterology*, 89:1084–1091, 1985.

[16] J. Cohen. The Earth Is Round - P-value less than 0.05. *American Psychologist*, 49(12):997–1003, 1994.

[17] Cook and Weisberg. Characterisations of an empirical influence function for detecting influential cases in linear regression. *Technometrics*, 22:495–508, 1980.

[18] D. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[19] D. R. Cox. Regression models and life tables. *J. R. Statist. Soc. B*, 34:187–220, 1972.

[20] A. Demaris. Explained variance in logistic regression a Monte Carlo study of proposed measures. *Sociological Methods and Research*, 77:329–342, 2002.

[21] N. R. Draper and H. Smith. *Applied Regression Analysis*. New York: John Wiley, 3 edition, 1998.

[22] D. Dunkler, S. Michiels, and M. Schemper. Gene expression profiling: Does it add predictive accuracy to clinical characteristics in cancer prognosis? *European Journal of Cancer*, 43:745–751, 2007.

[23] N. Ebrahimi and S. N. U. A. Kirmani. A characterisation of the proportional hazards model through a measure of discrimination between two residual life distributions. *Biometrika*, 83(1):233–235, 1996.

[24] M. Ezekiel. The sampling variability of linear and curvilinear regressions a first approximation to the reliability of the results secured by the graphic successive approximation method. *The Annals of Mathematical Statistics*, 1(4):275–333, 1930.

[25] C. Fan, S. Daniel, L. Wessels, B. Weigelt, D. Nuyten, A. B. Nobel, L. J. Vant Veer, C. M. Perou. Dore, A. Charlett, and J. D. Lewis. Concordance among gene-expression-based predictors for breast cancer. *New England Journal of Medicine*, 355:560–569, 2006.

[26] A. I. Fleishman. A method for simulating non-normal distributions. *Psychometrika*, 43:521–531, 1978.

[27] T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis.* Wiley and Sons, New York, 1991.

[28] L. S. Freedman and B. I. Graubard. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11:167–178, 1992.

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

[30] E. Graf. *Explained Variation Measures in Survival Analysis.* PhD thesis, University of Freiburg in German, 1998.

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

[32] E. Graf and M. Schumacher. An investigation on measures of explained variation in survival analysis. *The Statistician*, 44(4):497-507, 1995.

[33] S. Greenland. A lower bound for the correlation of exponentiated bivariate normal pairs. *American Statistician*, 50(2):163–164, 1996.

[34] J. W. Hardin and J. M. Hilbe. *Generalised Linear Models and Extensions.* Stata Press, Texas, 2 edition, 2007.

[35] F. E. Harrell. The phglm procedure. *SUGI Supplemental Library Users Guide*, 5:437–466, 1986.

[36] F. E. Harrell. *Regression Modeling Strategies.* Springer-Verlag, New York, 2001.

[37] F. E. Harrell and K. L. Lee. Verifying assumptions of the cox proportional hazards model. *In Proceedings of the Eleventh Annual SAS Users Group International Conference*, pages 823–828, 1986.

[38] R. V. Hartley. Transmission of information. *Bell Systems Technical Journal*, 7:535–563, 1928.

[39] W. W. Hauck, S. Anderson, and S. M. Marcus. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials*, 19:249–256, 1998.

[40] H. Heinzl. Using sas to calculate the kent and O'Quigley measure of dependence for cox proportional hazards regression model. *Computer Methods and Programs in Biomedicine*, 63:71–76, 2000.

[41] I. S. Helland. On the interpretation and use of R2 in regression analysis. *Biometrics*, 43(1):61–69, 1987.

[42] R. Henderson. Problems and prediction in survival data. *Statistics in Medicine*, 14:161–184, 1995.

[43] R. Henderson, M. Jones, and J. Stare. Accuracy of point prediction in survival analysis. *Statistics in Medicine*, 20:3083–3096, 2001.

[44] R. Henderson and P. Oman. Robust estimation in cox regression model. *Scandinavian Journal of Statistics*, 20:195–212, 1993.

[45] J. Herson. The use of surrogate end points in clinical trials. *Statistics in Medicine*, 8:403–404, 1989.

[46] Y. Huang and N. R. Draper. Transformations, regression geometry and r2. *Computational Statistics and Data Analysis*, 42(4):647–664, 2003.

[47] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

[48] M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics.* Macmillan, New York, 1979.

[49] J. Kent and J. O'Quigley. Measures of dependence for censored survival data. *Biometrika*, 75(3):525–534, 1988.

[50] J. T. Kent. Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173, 1983.

[51] J. T. Kent. The underlying structure of nonnested hypothesis tests. *Biometrika*, 73(2):333–343, 1986.

[52] R. E. Kirk. Effect magnitude: A different focus. *Journal of Statistical Planning and Inference*, 137:1634–1646, 2007.

[53] E. L. Korn and R. Simon. Measures of explained variation for survival data. *Statistics in Medicine*, 9:487–503, 1990.

[54] E. L. Korn and R. Simon. Explained residual variation, explained risk, and goodness of fit. *The American Statistician*, 45(3):201–206, 1991.

[55] S. Kullback. *Information Theory and Statistics.* New York Dover, 1951.

[56] T. O. Kvalseth. Cautionary note about R2. *The American Statistician*, 39(4):279–285, 1985.

[57] S. W. Lagakos and D. A. Schoenfeld. Properties of proportional hazards score tests under misspecified regression models. *Biometrics*, 40(4):1037–1048, 1984.

[58] A. Laupacis, N. Sekar, and I. G. Stiell. Clinical prediction rules. *JAMA*, 277(6):488–494, 1997.

[59] J. F. Lawless. *Statistical Models and Methods for lifetime Data.* Wiley and Sons, New York, 1982.

[60] J. F. Lawless. *Statistical Models and Methods for lifetime Data.* Wiley and Sons, New York, 2 edition, 2002.

[61] L.M. Leemis. Variate generation for accelerated life and proportional hazards models. *Operations Research*, 35:892–894, 1987.

[62] L.M. Leemis, L. H. Shih, and K Reynertson. Variate generation for accelerated life and proportional hazards models. *Statistics and Probability Letters*, 10:335–339, 1990.

[63] L. Lininger, M. H. Gail, S. B. Green, and D. P. Byar. Comparison of four tests for equality of survival curves in the presence of stratification and censoring. *Biometrika*, 66(3):419–428, 1979.

[64] C. L. Link. Confidence intervals for the survival function using Cox's proportional hazards model with covariates. *Biometrics*, 40:601–610, 1984.

[65] D. A. Ludwig. Null hypothesis significance testing - a review of an old and continuing controversy. *Psychological Methods*, 5(2):241–301, 2000.

[66] D. A. Ludwig. Use and misuse of p-Values in designed and observational Studies - guide for researchers and reviewers. *Aviation, Space, and Environmental Medicine*, 76(7):675–680, 2005.

[67] G. S. Maddala. *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, Cambridge UK, 1983.

[68] L. Magee. R2 measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*, 44(3):250–253, 1990.

[69] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis.* John Willey, New York, 3 edition, 2001.

[70] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics.* McGraw-Hill, Singapore, 3 edition, 1974.

[71] N. J. D. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692, 1991.

[72] A. Nardi and M. Schemper. New residuals for cox regression and their application to outlier screening. *Biometrics*, 55:523–529, 1999.

[73] B. M. Ogles, K. M. Lunnen, and K. Bonesteel. Clinical significance - History, application, and current practice. *Clinical Psychology Review*, 21(3):421–446, 2001.

[74] J. O'Quigley. *Proportional Hazards Regression.* Springer, New York, 2008.

[75] J. O'Quigley and P. Flandre. Predictive capability of proportional hazards regression. *Proceedings of the National Academy of Sciences of the United States of America*, 91:2310–2314, 1994.

[76] J. O'Quigley and P. Flandre. Quantification of the prentice criteria for surrogate endpoints. *Biometrics*, 62:297–300, 2006.

[77] J. O'Quigley, P. Flandre, and E. Reiner. Large sample theory for schemper's measure of explained variation in the cox regression model. *The Statistician*, 48(1):53–62, 1999.

[78] J. O'Quigley and R. Xu. *Handbook of Statistics in Clinical Oncology*, chapter 19, pages 397–410. Marcel Dekker, New York, 2001.

[79] J. O'Quigley and R. Xu. *Handbook of Statistics in Clinical Oncology*, chapter 19, pages 347–364. Marcel Dekker, New York, 2006.

[80] J. O'Quigley, R. Xu, and J. Stare. Explained randomness in proportional hazards models. *Statistics in Medicine*, 24:479–489, 2005.

[81] K Pearson. Regression, heredity, and fanmixia. *Phil. Trms. Roy. Soc., Series A*, clxxxvii:253, 1896.

[82] A. N. Pettitt and I. Bin-Daud. Case-weighted measures of influence for proportional hazards regression. *Applied Statistics*, 38(1):51–67, 1989.

[83] R. L. Prentice. Surrogate end points in clinical trials: definition and operating criteria. *Statistics in Medicine*, 8:431–440, 1989.

[84] A. Ritchie, G. Griffiths, and M. Parmar. Interferon-alfa and survival in metastatic renal carcinoma: early results of a randomised controlled trial. *Lancet*, 353:17–17, 1999.

[85] A. Rosenwald, G. Wright, W. C. Chan, and et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346:1937–1947, 2002.

[86] S. Rosthoj and N. Keiding. Explained variatioan and predictive accuracy in general parametric statistical models: The role of model misspecification. *Lifetime Data Analysis*, 10:461–472, 2004.

[87] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* New York: John Wiley, 1 edition, 1987.

[88] P. Royston. Explained variation for survival models. *The Stata Journal*, 6(1):1–14, 2006.

[89] P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates parsimonious parametric modelling. *Applied Statistics*, 43(3):429–467, 2006.

[90] P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141, 2006.

[91] P. Royston, J. Atzpodien, and M. Reitz. An approach to estimating prognosis using fractional polynomials in metastatic renal carcinoma. *British Journal of Cancer*, 94:1785–1788, 2006.

[92] P. Royston and M. K. B. Parmar. Flexible parametric models for censored survival data with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21:2175–2197, 2002.

[93] P. Royston and W. Sauerbrei. A new measure of prognostic separation in survival data. *Statistics in Medicine*, 23:723–748, 2004.

[94] W. Sauerbrei and P. Royston. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society (Series A)*, 162:71–94, 1999. Corrigendum: *Journal of the Royal Statistical Society (Series A)*, 165:399–400, 2002.

[95] M. Schemper. The explained variation in proportional hazards regression. *Biometrika*, 77(1):216–218, 1990.

[96] M. Schemper. Amendments and corrections: The explained variation in proportional hazards regression. *Biometrika*, 81(3):631, 1994.

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

[98] M. Schemper and A. Kaider. A new approach to estimate correlation coefficient in the presence of censoring and proportional hazards. *Computational Statistics and Data Analysis*, 23:467–476, 1997.

[99] M. Schemper and J. Stare. Explained variation in survival analysis. *Statistics in Medicine*, 15:1999–2012, 1996.

[100] C. Schmoor and M. Schumacher. Effects of covariate omission and categorization when analysing randomised trials with the cox model. *Statistics in Medicine*, 16:225–237, 1997.

[101] D.A Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrica*, 69:239–241, 1982.

[102] M. Schumacher, G. Bastert, H. Bojar, K. Hubner, M. Olshewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. L. A. Newmann, and H. F. Rauschecker. Randomised

2*2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12:2086–2093, 1994.

[103] A. Scott and C. Wild. Transformation and R2. *The American Statistician*, 45(2):127–129, 1991.

[104] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[105] R. Simon. Importance of prognostic factors in cancer clinical trials. *Cancer Treatment Reports*, 68:185–192, 1984.

[106] J. M. Smith, C. J. Dore, A. Charlett, and J. D. Lewis. A randomised trial of biofilm dressing for venous leg ulcer. *Phlebology*, 7:108–113, 1992.

[107] E. S. Soofi. Capturing the intangible concept of information. *Journal of the American Statistical Association*, 89(428):1243–1254, 1994.

[108] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[109] J. Stare. *Measures of explained variation in survival analysis.* PhD thesis, Medical Faculty, University of Ljubljana, Slovenia, 1994.

[110] J. W. Tukey. *Exploratory Data Analysis.* Addison-Wesley Publishing Company, Philippines, 1 edition, 1977.

[111] W. Vandaele. Wald, likelihood ratio, and lagrange multiplier tests as an F test. *Economics Letters*, 8:361–365, 1981.

[112] L. J. Vant Veer, H. Dai, M. J. Vijver, and et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[113] P. J. M. Verweij and H. C. Van Houwelingen. Cross-validation in survival analysis. *Statistics in Medicine*, 12:2305–2314, 1993.

[114] R. R. Wilcox and J. Muska. Measuring effect size: A non-parametric analogue of w2. *British Journal of Mathematical and Statistical Psychology*, 52:93–110, 1999.

[115] R. Xu. *Inference for the proportional Hazards Model.* PhD thesis, University of California, San Diego, 1996.

[116] R. Xu and J. O'Quigley. A measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics*, 12:83–107, 1999.