INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Impact of Data Dependent Model Selection on Inference

Mamun Mahmud

A Thesis

In

The Department

of

Mathematics and Statistics

Presented in Partial Fulfilment of the Requirements
For the Degree of Master of Science at
Concordia University
Montreal, Quebec, Canada

August 2000

Impact of Data Dependent Model Selection on Inference

Mamun Mahmud

Abstract

Many continuous variables may have non-linear effects on the outcomes of interest. For example, the effect of serum cholesterol on the logit of the probability of coronary heart disease mortality is stronger in the low to moderate range of observed cholesterol values than in the upper tail of their distribution (Abrahamowicz *et al.*, 1997). If so, it is often not obvious how to represent the functional relationship between the covariate and the outcome. One common approach is to estimate different models, each using a different function of the covariate of interest (typically corresponding to different parametric transformations), and then to select the one that fits the data best. However, this approach can be considered as a specific case of the general problem of data-dependent model selection. As the model selection uncertainty is typically not accounted for, such approach is likely to induce some bias at the step of statistical inference.

In this thesis, I consider the problem of accounting for model uncertainty in a parametric regression model with focus on the uncertainty involved in selection of the *optimal* transformation of a continuous predictor in the Cox proportional hazards model (Cox, 1972). I use the minimum AIC approach to select *a posteriori* the optimal transformation of a continuous predictor. First, I review literature on criteria and methods for selecting the *"best-fitting"* model based on the results

obtained from a sample, in Chapter 1. Then, in Chapter 2, I discuss the general problem of model selection uncertainty on inference and summarize research in this area. Next, I evaluate the impact of the data-dependent model selection approach on *type I error* rate through simulations. In simulations, I generate data, assuming both linear and non-linear dependence of hazard on a continuous covariate as well as no association. The generated data are then used to estimate a series of models with various functional form, to assess the impact of model selection on *type I error* and on statistical power. Some of the above methodological problems are then illustrated in the analysis of a real-life dataset including several cardiovascular risk factors.

Acknowledgements

I am greatly indebted to my supervisor Dr. Michal Abrahamowicz of McGill University for his tremendous help and encouragement at all stages of writing this thesis. He provided me with invaluable statistical guidance and endless support in the form of constructive discussions and necessary corrections. I am sure that without his help, encouragement, and financial support from his NSERC grant, I could never have completed this thesis.

There are many other people who helped me in completing this work. I would like to acknowledge my indebtness to my co-supervisor Dr. Y.P. Chaubey for his valuable suggestions and encouragement. I would like to thank Ms. Roxane du Berger for her patience in helping me to familiarize myself with the S-Plus environment. My special thanks are due to Ms. Karen Leffondré who read the entire manuscript carefully, introduced many corrections and helped me to reorganize the material. Without her help I could never have produced the thesis in the present form.

My thank go to the staff of the Clinical Epidemiology Division, particularly to Ms. Jennifer Gardener and Ms. Mary for their secretarial support.

I wish to thank the Department of Mathematics and Statistics of Concordia University, in particular the faculty, for their invaluable instructions during my period of study, and the staff for their timely support and assistance.

Finally, I wish to thank my family for their continuous encouragement and understanding during the period of my studies at Concordia University.

Contents

List of Figures x

List of Tables xi

1 Review of Criteria and Methods for Model Selection **1**

1.1 Introduction 1

1.2 Overview of the Thesis 3

1.3 The Importance of Model Selection in the Regression Analysis 4

1.4 Selection Criteria 9

1.4.1 Cross-Validation 10

1.4.2 Criteria Specific for Linear Regression: Adjusted R^2 and Mallow's C_p 11

1.4.3 Akaike Information Criterion (AIC) 12

1.4.4 Quasi-Likelihood and Takeuchi's Information Criterion (TIC) 15

1.4.5 Bayesian Approaches 17

1.4.6 Other Criteria 18

1.5 Summary and Discussion 19

2 Model Selection Uncertainty **22**

2.1 Introduction 22

2.2 Confidence Subset of Models 26

2.2.1 Method Based on Akaike Weight (w_i) 27

2.2.2 Method Based on Cut-off Δ_i 32

2.2.3 Method Based on Relative Likelihood 33

2.2.4 Summary 34

2.3 Methods of Assessing Model Selection Uncertainty 34

2.3.1 General Comments on Bootstrap 36

2.3.2 AIC Differences (Δ_i), Model Selection Probability, and the Bootstrap 37

2.3.3 Monte Carlo 39

2.4 Uncertainty in Parameter Estimates Associated With Model Selection 39

2.4.1 Including Model Selection Uncertainty in Estimating Sampling Variance 42

2.5 Variable Selection in Multiple Regression and Model Uncertainty . . . 45

2.6 Model Selection In Survival Analysis 49

2.7 Summary 51

3 Model Selection Problems in the Context of Choosing the Optimal Transformation of a Continuous Covariate in Cox's Regression **54**

3.1 Introduction 54

3.1.1 Representation of the Effect of a Continuous Covariate as a Specific Area Where Data-Dependent Model Selection is often Employed 55

3.1.2 Overview of Conventional Approaches to Modelling the Effects of Continuous Covariates in Epidemiology 58

3.1.3 Overview of Conventional Regression Models For Predicted-Response Relationships 60

3.1.4 Logistic Regression Model 62

3.1.5 Proportional Hazards Regression Model 63

3.2 Design of the Simulation Study 68

3.2.1 Data Generation Procedure 68

3.2.2 Considered Configurations 70

3.2.3 Details of the Data Generation Algorithm 74

3.2.4 Data Analytical Procedures 77

4 Results **81**

4.1 Results of the Simulations Study 81

4.1.1 Assessing the Impact of Model Selection on Type I Error Rate 81

4.1.2 Preliminary Evaluation of Two Simple Model Averaging Approaches to Testing the Hypothesis of No Association 88

4.1.3 Comparison of the Statistical Power of Different Testing Procedures 89

4.2 Real Life Illustration 94

4.2.1 Data Description 94

4.2.2 Parametric Modelling of Individual Risk Factors 95

5 Conclusions **102**

Bibliography **104**

A S-PLUS Code For The Simulation Study **120**

A.1 Simulation for the Unselected Models 120

A.2 Simulation for the Best AIC models 123

A.3 Additional Code for the calculation of the Empirical Power 126

A.4 Drawing the Histograms 128

List of Figures

1.1 The Principle of Parsimony 5

3.1 Functions Representing the Linear and Non Linear Dose-Response Relationship 72

4.1 Figures Showing Systematic Bias in the Distribution of p-values . . . 86

4.2 Plot of Body Mass Index (BMI) Against Log of Hazard Ratio 97

List of Tables

2.1 Relationship between AIC differences (Δ_i), Relative Likelihood $\exp[-\frac{1}{2}\Delta_i]$ and Normalized Akaike weights for 7 hypothetical models 30

2.2 Coverage Rates for Confidence Intervals, Conditional on Selected Model with Nominal Rate $(1-\alpha)$, $n = 20$, Correct Model, $p_0 = 3$ . . 47

4.1 Observed Type I Error Rates of The LR Test For The Nominal Size $\alpha = 0.05$ and Proportion of Samples Where a Given Model Was Selected as the Best AIC Model 83

4.2 Proportion of Simulated Samples, Where H_0 was Rejected at $\alpha = 0.05$, When There is No Covariate Effect: Effect of Model Averaging on Type I Error 89

4.3 Comparison of the Empirical Power of the Two Testing Procedures in the Case of a Linear Association Between Covariate and log Hazard 91

4.4 Comparison of the Empirical Power of the four Testing Procedures For Different Sample Sizes and Forms of Dose-Response Curve ($\alpha = 0.05$) 92

4.5 Results For the Cox PH Model With Different Functional Forms of TC 99

4.6 Results For the Cox Model With Different Functional Forms of SBP . 99

4.7 Results of the Cox Model With Different Functional Forms of BMI . 100

4.8 Results of the Cox Model With Different Functional Forms of TC/HDL 100

4.9 Results of the Cox Model With Different Functional Form of Age . . 101

Chapter 1

Review of Criteria and Methods for Model Selection

1.1 Introduction

A large part of statistical research focuses on modelling data. On the other hand, one of the most important problems confronting an investigator in statistical modelling is the choice of an appropriate model to characterize the underlying data. This determination can often be facilitated through the use of an information-theoretic criterion, which judges the propriety of a fitted model by assessing whether it offers an optimal balance between "*goodness of fit*" and parsimony.

There is substantial statistical literature discussing *model selection*, which is the practice of selecting a model to fit the data at hand. A typical approach to data analysis involves three main steps. First, postulate the class of competing models using a modern computing power, then typically select the single model which is *best* according to some predetermined criteria, and finally, make inference as if the selected model is the *true* model. It is well known that, when a model is formulated

and fitted to the same set of data, inferences made from it will be biased and over-optimistic when they ignore the data analytic actions which preceded the inference (Chatfield, 1995).

In epidemiological studies, a common problem is to determine the specific mathematical relationship between the value of a continuous covariate and the risk of disease occurrence. Evaluation of various statistical methods to describe accurately associations between exposures and disease are constantly being explored. As complexity of relationships investigated in empirical sciences increases, we will need to rely more on model selection criterion. However, different choices of a model selection criterion may produce different models. Moreover, even with the same criterion, the "*optimal*" model may vary from sample to sample. Thus, while model selection may be desirable or even necessary in many applications, there is uncertainty about which model to use for making valid inference regarding model parameters. Therefore, it is important to explore the implications of model selection on traditional inference about model parameters.

However, there has been relatively little exploration of the practical implication of their effects. Given the analytical methods are generally not available to study the effects of data-dependent procedures, a variety of computational methods have been tried. *Simulation* methods are one obvious avenue when the model selection procedure is simple and clearly defined. In this thesis, I will evaluate the impact of data-dependent model selection on statistical inference through *simulation studies*

as well as through an empirical study.

1.2 Overview of the Thesis

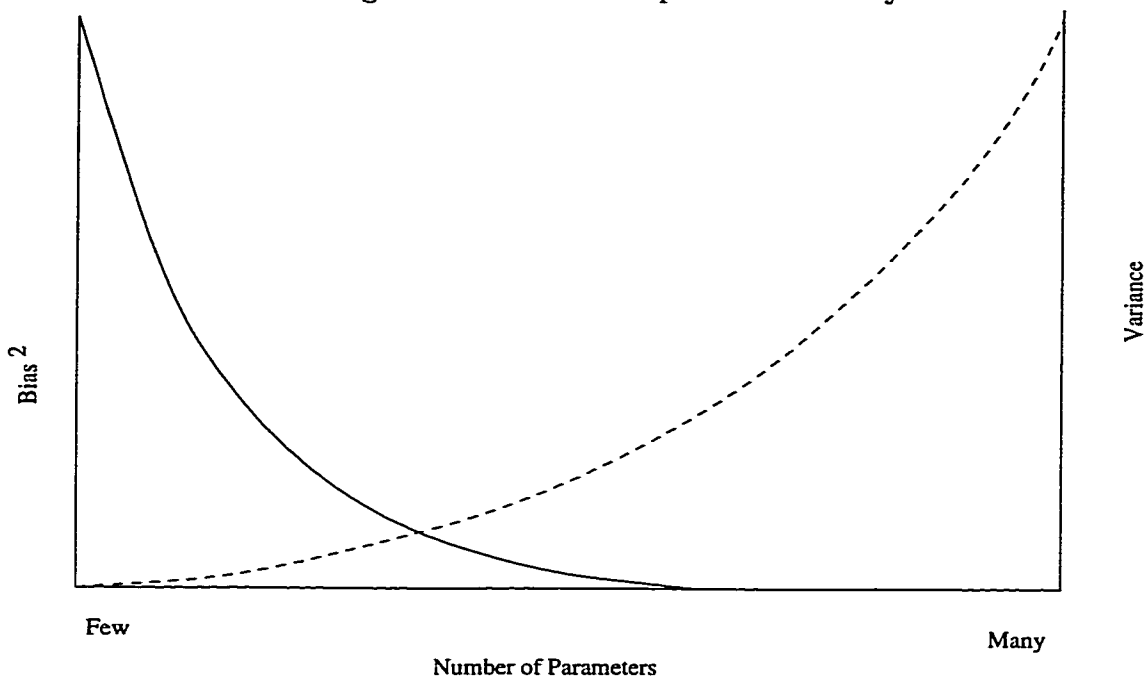
In this thesis, I discuss the problem of model uncertainty in the context of censored survival model, that evolves due to data dependent *a posteriori* model selection and its probable impact on subsequent inference. In Chapter 1, I review the literature on criteria and methods for model selection with a look on its importance in regression analysis. Chapter 2 presents a detailed overview of the the issues related to model selection uncertainty that arises due to the model building process. Chapter 3 consists of two parts. First part discusses the model selection problem in the context of choosing the *optimal* transformation of a continuous covariate in Cox proportional hazards regression; second part presents the simulation studies I carried out to evaluate the impact of model selection. Chapter 4 mainly speaks about the analysis of the simulated and the real life data-sets. Finally, Chapter 5 is a discussion of the conclusions I reach in this thesis. The S-Plus code I created for this thesis is presented in the appendix **A**.

1.3 The Importance of Model Selection in the Regression Analysis

Statistical models are useful in the empirical sciences for understanding the inherent structure of data, description of the data and for making inference. In recent years, the problem of *model selection* recognized in literature can be stated as follows: Given a data set, how do we choose the "*best approximating*" model among a class of competing models, possibly with different numbers of parameters, by a suitable model selection criterion? (Bozdogan, 1987). Success in the analysis of real data and the resulting inference often depend importantly on the choice of the *"approximating model"*. In the biological sciences, such analysis is required to be based on a parsimonious model that provides an accurate approximation to the structural information in the data at hand. This should not be viewed as searching for the *"true model"* but rather coming up with the "*best approximating*" model. As such model selection is essentially concerned with the *"art of approximation"* (Akaike, 1974).

Box and Jenkins (1970) suggested that the principle of parsimony should lead to a model with "the smallest possible number of parameters for adequate representation of the data". Parsimony is the concept that a model should be as simple as possible with respect to the number of included variables, model structure and number of parameters. Parsimony is a desired characteristic of a model used for inference and it is usually operationalized as a suitable trade-off between bias and variance

Figure 1.1: The Principle of Parsimony



of parameter estimators. In general, magnitude of the bias decreases and variance increases as the dimension of the model, i.e. the number of estimated parameters (K), increases (Figure 1.1).

All model selection methods are based to some extent on the principle of parsimony (Breiman, 1992; Zhang, 1994). In understanding the utility of an approximating model for a given data set, it is convenient to consider two undesirable possibilities: under-fitted and over-fitted models. The terms under and over-fitted models are used in relation to a "*best approximating model*". An under-fitted model would ignore some important replicable structure in the data and thus fail to identify effects that were actually supported by the data. As a consequence, bias in the parameter estimation is often substantial and the sampling variance is typically

underestimated, both factors resulting in poor coverage probability (Burnham and Anderson, 1998). On the other hand, over-fitted models, as judged against a best approximating model, are often free of bias in the parameter estimators, but provide estimated sampling variances for the same that are needlessly large (i.e. the precision of the estimator is poor, relative to what could have been accomplished with a more parsimonious model). Shibata (1989) argued convincingly that under-fitted models pose a more serious problem in data analysis and inference than over-fitted models. In fact, the best approximating model is achieved by properly balancing the errors of under-fitting and over-fitting. The proper balance is achieved when bias and variance are controlled to achieve confidence interval coverage at approximately the nominal level and when interval width is at a minimum. Achieved confidence interval coverage is a convenient index of whether the accuracy of both parameter estimators and measures of precision are adequate. Proper model selection procedures attempt to identify a model in which the error of approximation and the error due to random fluctuations are well balanced (Shibata, 1983, 1989). Some model selection methods, e.g. Bayesian Information Criterion (BIC) (Schwarz, 1978) are parsimonious but tend to select models that may be too simple (i.e. under-fitted); thus bias is large, precision is over-estimated and achieved confidence interval coverage is well below the nominal level. Such instances are not satisfactory for inference because apparent high precision of the estimates is misleading, given their substantial bias (Burnham and Anderson, 1998).

The impact of model selection on inference has most often been viewed in the context of hypothesis testing. Sequential testing has most often been employed, based on either step up (forward) or step down (backward) methods (Burnham and Anderson, 1998). Procedures using model selection testing schemes are based on subjective α levels (commonly 0.05 or 0.01); however Rawlings (1988) recommends $\alpha = 0.15$ in the context of stepwise regression. The multiple testing problem is serious if many tests are to be made and the tests are not independent (Westfall and Young, 1993).

A substantial limitation in the use of hypothesis testing for model selection is that traditional likelihood ratio tests (LRT) are defined only for nested models; so tests between models that are not nested are problematic. Some authors argue that hypothesis testing is a poor basis for model selection (Akaike, 1974; Sclove, 1994). Akaike (1974) noted, "The use of a fixed level of significance for the comparison of models with various numbers of parameters is wrong, since it does not take into account the increase of the variability of the estimates when the number of parameters is increased". In fact, the significance level should be related to sample size and the degrees of freedom, if the hypothesis testing is to be somehow used as a basis for model selection (Akaike, 1974). However, in the hypothesis testing approach, the α-level is usually kept fixed regardless of sample size or degrees of freedom. This practice of keeping α-level constant implies asymptotically inconsistent results in hypothesis testing. For example, if the null hypothesis is true and α is fixed (at, say,

0.05), then even as degrees of freedom approach ∞, we still have a 0.05 probability of rejecting the null hypothesis, even with an almost infinite sample size. Yet, to be consistent, the statistical procedures in this simple context should converge on truth with probability 1 as the sample size (n) tends to infinity.

If goodness of fit tests can be computed for all alternative models even if some are not nested within others, then one could use the model with the fewest number of parameters that *"fits"* the best i.e. yields $p > 0.05$ or 0.10 for the test of the null hypothesis of adequate fit. However, increasingly better fits can often be achieved by using models with more parameters and this can make the arbitrary choice of α very critical. A large α level leads to over-fitted models and their resulting problems. In fact, there is no theory to suggest that this approach will lead to selected models with good inferential properties (i.e. an adequate bias vs. variance trade-off or acceptable coverage and/or width of the confidence interval). In addition, other problems may be encountered such as over- or under-dispersion and low power (Burnham and Anderson, 1998).

Truth in the biological sciences and medicine is inherently complicated and we can not hope to find the "*true*" model from the analysis of a finite amount of data. Thus, inference about truth must be based on a good "*approximating model*". Likelihood and least square methods provide a rigorous inference, if the model structure is given. However, in most real-life applications, the model is not *given*. Thus the critical issue is "*what model to use*"? This is the model selection problem. The em-

phasis then shifts to the careful *a priori* definition of a set of candidate models. Information-theoretic approach provides a simple way to select a "*best*" approximating model from the candidate models. In fact, information theory based on Kullback-Liebler (K-L) information provides a sound theoretical basis for model selection, while likelihood ratio test (LRT) doesn't.

1.4 Selection Criteria

As far as a model yields a *good* approximation, a simpler model is better than a complex one both for understanding the underlying phenomenon and for various applications. The principle is the same for selecting a statistical model. Complexity of a model is restricted both by the size of observation and by the number of parameters included in the model. Needless to say, complete specification might be possible if an infinite number of observations were available for a quite simple system. Otherwise, a practical procedure is, starting from a simple model, to increase the complexity until a trade-off between the error of approximation and the error due to random fluctuations is obtained (Shibata, 1989). To do this systematically, a convenient way is to introduce a formal *criterion* to compare models. In the next sections, I shall discuss various criteria based mainly on information theory as well as some re-sampling procedures.

1.4.1 Cross-Validation

Cross-validation has been suggested and well studied as a basis for model selection (Mosteller and Tukey, 1968; Stone, 1974, 1977; Geisser, 1975). Different procedures are used for Cross-Validation. In the simplest case, the data are divided into two subsets, the first subset is used for model fitting and the second is used for model validation (sometimes the second subset has only one observation). Then a new subset is selected, and this whole process is repeated a large number of times. Also then some criterion is chosen such as minimum squared prediction error (MSPE), as a basis for model selection. There are several variations on this theme, e.g. leaving out one (or more) observation at a time, a model is fitted to the remaining points and used to predict the deleted points. Thus, within-sample prediction errors provide an assessment of prediction quality of the models. The sum of squares of these errors can also be used to make a choice between different models. There are several variations of this approach (Craven and Wahba, 1979; Burman, 1989; Shao and Tu, 1995; Zhang, 1993a; Hjorth, 1994). These methods are quite computer intensive and tend to be impractical if more than about $15 - 20$ models must be evaluated or if sample size is large. Still, cross-validation offers an interesting alternative for model selection.

1.4.2 Criteria Specific for Linear Regression: Adjusted R^2 and Mallow's C_p

The selection of models using the adjusted R^2 statistic and Mallow's C_p are related for least squares problems (Saber, 1977). The adjusted coefficient of multiple determination $\bar{R}_p^2$ has been used in model selection in a least square setting and is defined as

$$\bar{R}_p^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-p)},$$

where R^2 is the usual coefficient of multiple determination and p is the number of parameters in the model of n observations (Draper and Smith 1981). While R^2 automatically increases as a new parameter is introduced, $\bar{R}_p^2$ does not and can be compared for models containing different numbers of parameters. The models giving the highest $\bar{R}_p^2$, closest to one, are taken as the best fitting. Thus under this method, one selects the model in which the adjusted statistics is largest.

Mallow's C_p statistic (Mallows, 1973, 1995) is used in least square regression with normal residuals and a constant variance. In this special case, it provides a ranking of the candidate models, which is the same as the ranking under Akaike Information Criterion (AIC, see the next section) (Atilgan, 1996). Mallow's C_p can be formulated as a direct adjustment of the residual sum of squares (RSS) based on p variables in the model. It takes the form:

$$C_p = \frac{RSS}{\hat{\sigma}^2} + 2p - n,$$

where $\hat{\sigma}^2$ is an estimate of the underlying variance σ^2, usually based on fitting all

the regressors. For a correct model $E(C_p) = p$, hence a large deviation of C_p from the line $C_p = p$ suggests a wrong model and consequent bias in the fitted model.

1.4.3 Akaike Information Criterion (AIC)

The first information theoretic criteria to gain wide-spread acceptance as a model selection tool was the *AIC* (Akaike, 1973, 1974, 1977, 1978a, 1978b, 1981a, and 1981b). Akaike developed this information-theoretic, or entropic *AIC* criterion, which takes model complexity into account, for the identification of an optimal parsimonious model from a class of competing models. This criterion, used in the non nested case, has been based on the likelihood function with best estimates of the parameters and an adjustment for the number of parameters. It attempts to balance the need for a model which fits the data very well to that of having a simple model with few parameters. Akaike's information-theoretic approach has led to a number of alternative methods having desirable properties for the selection of best approximating models in practice (see Burnham and Anderson, 1998).

Kullback and Liebler (1951) derived an information measure that happened to be negative of Boltzman's entropy, which is nowadays known as the Kullback-Liebler (K-L) distance. The K-L distance (see Section 2.2 for details) can be conceptualized as a directed "*distance*" between two models, say "*true*" model f and candidate g (Kullback, 1959). Akaike proposed the use of this Kullback-Liebler (K-L) distance as a fundamental basis for model selection. He found a relation between the relative K-L distance and Fishers maximized log likelihood function and on this basis proposed

the famous criterion

$$AIC = -2\log L(\hat{\theta}|y) + 2K, \tag{1.1}$$

where $L(\theta|y)$ is the likelihood of the parameters θ given the data y, K is the number of parameters used in the model, and $\hat{\theta}$ is the MLE of θ. The first term in (1.1) is a measure of inaccuracy, badness of fit, or bias when maximum likelihood estimators of the parameters of the model are used. The second term, on the other hand, is a measure of complexity or the penalty for the increased unreliability in the first term, which depends upon the number of parameters used to fit the data. So, the addition of twice the number of parameters is a penalty to correct for the expected reduction in estimation bias with increasing number of parameters (Bozdogan, 1987). This criterion will decrease as variables are added to the model. At some point, the criterion will increase and this is a signal that the added variables are unnecessary.

Thus, when there are several competing models, the parameters within the models are estimated by the method of maximum likelihood, and the value of the $AIC's$ are computed and compared to find a model with the minimum AIC value. This procedure is called the minimum AIC procedure and the model with the minimum AIC is called the minimum AIC estimate (MAICE) and is chosen to be the best model (Bozdogan, 1987). Therefore, the best model is the one with least complexity, or equivalently, the highest information gain. In applying AIC, the emphasis is on comparing the goodness of fit of various models with an allowance made for parsimony.

Although AIC is an asymptotically unbiased estimate of relative K-L distance, it is not true without assuming that true f is in the set of candidate models. To compare models, it is recommended to compute differences (rather than the actual AIC values),

$$\Delta_i = AIC_i - \min AIC, \tag{1.2}$$

over all candidate models i ($i = 1, 2, ..., R$, R being the number of candidate models) in the set. Here Δ_i is an estimate of the difference between the K-L distance between a given model i and the AIC optimal model. These Δ_i values are easy to interpret and allow a quick comparison and ranking of candidate models and are also useful in computing Akaike weights (see Section 2.4 for details). The larger is Δ_i, the less plausible is the fitted model i. Thus, model selection becomes a simple function minimization, where AIC is the criterion to be minimized.

However, AIC may perform poorly if there are too many parameters in relation to the size of the sample (Suguira, 1978; Sakamoto *et al.*, 1986). Suguira (1978) derived a second-order variant of AIC that he called $cAIC$:

$$cAIC = -2 \log L(\hat{\theta}) + K[\log n + 1]. \tag{1.3}$$

where, $L(\theta)$ is the likelihood of parameter θ. Note that $cAIC(k)$ is similar to the Schwarz's (1978) criterion of $K \log n$ (i.e. BIC, see Section 1.2.5), and that the term $[K \log n + K]$ has the effect of increasing the "penalty term". Consequently, the minimization of $cAIC$ leads, in general, to lower dimensional models than those obtained by minimizing AIC.

Hurvich and Tsai (1989) further studied this small sample (second-order) bias adjustment which led to a criterion that is called AIC_c:

$$\begin{aligned} AICc &= -2\log L(\hat{\theta}) + 2K(\frac{n}{n-K-1}) \\ &= AIC + \frac{2K(K+1)}{n-K-1}, \end{aligned} \tag{1.4}$$

If n is large with respect to K (number of parameters), then the second order correction is negligible and AIC should perform as well as AIC_c. Burnham *et al.* (1994) suggested that AIC_c has to be used when the ratio $\frac{n}{K}$ is small (say < 40). If the ratio $\frac{n}{K}$ is sufficiently large, then AIC and AIC_c are similar and will tend to select the same model. However unless the sample size is large with respect to the number of estimated parameters, use of AIC_c is recommended.

AIC selection is objective and represents a very different paradigm to that of hypothesis testing and is free from the arbitrary α levels, the multiple testing problem, and the fact that some candidate models might not be nested. AIC allows a ranking of models and the identification of models that are nearly equally useful, versus those that are poor explanations for the data at hand. The AIC is reminiscent of the adjusted R^2 in least-squares regression, in that both are attempting to adjust the fit of the model for the number of parameters included.

1.4.4 Quasi-Likelihood and Takeuchi's Information Criterion (TIC)

In the case of over-dispersion found in count data (i.e. when sampling variance exceeds theoretical model-based variance), one needs to model the over-dispersion

and then use generalized likelihood inference methods. Quasi-likelihood (Wedderburn, 1974) theory is a basis for the analysis of over-dispersed data (Williams, 1982; McCullagh and Pregibone, 1985; Moore, 1987; McCullagh and Nelder, 1989). Principles of quasi-likelihood suggest simple modifications to AIC and $AICc$. According to Lebreton (1992):

$$QAIC = -\left\{2 \log L(\hat{\theta})/\hat{c}\right\} + 2K, \tag{1.5}$$

and

$$\begin{aligned} QAICc &= -\left\{2 \log L(\hat{\theta})/\hat{c}\right\} + 2K + \frac{2K(K+1)}{n-K-1} \\ &= QAIC + \frac{2K(K+1)}{n-K-1}, \end{aligned} \tag{1.6}$$

where $\hat{c}$ is the variance inflation factor, estimated from goodness of fit of χ^2 statistic. When no over-dispersion exists ($\hat{c} = 1$), then formulae for $QAIC$ and $QAIC_c$ reduce to AIC and AIC_c, respectively.

TIC (Takeuchi's information criterion) allows for substantial model misspecification, i.e. applies in situations where perhaps none of the candidate models approximates well the "*true*" f. In this case there is a more general bias adjustment term to allow $-2 \log L(\hat{\theta})$ to be adjusted to be an asymptotically unbiased estimate of relative K-L, thus

$$TIC = -2 \log L(\hat{\theta}) + 2\, tr[J(\hat{\theta}) I(\hat{\theta})^{-1}], \tag{1.7}$$

where the matrices $J(\theta)$ and $I(\theta)$ involve first and second mixed partial derivatives of the log likelihood function, and 'tr' denotes the trace of the matrix.

AIC is equivalent to TIC only when $tr[J(\hat{\theta})I(\hat{\theta})^{-1}] \equiv K$. The benefit of TIC is that one achieves asymptotic unbiasedness for K-L model selection. One might consider always using TIC as this reduces the concern about the adequacy of the set of candidate models. If over-dispersion is found in count data, then the log-likelihood could be divided by an estimated variance inflation factor, giving a criterion called $QTIC$ (Burnham and Anderson, 1998).

Linhart and Zucchini (1986) proposed a further generalization and Amari (1993) proposed a network information criterion (NIC) potentially useful in training samples in neural network models. Shibata (1989) developed a complicated criterion, based on the theory of penalized likelihoods and called it RIC (regularized information criterion).

1.4.5 Bayesian Approaches

Bayesian researchers have taken somewhat different approaches and assumptions and have proposed several alternative methods for model selection. Some of these are difficult to implement and very computer intensive (Laud and Ibrahim, 1995; Carlin and Chib, 1995).

Several criteria have been developed based on the assumption that an exactly "*true model*" exists among the candidate models being considered and the model selection goal is to select the true model. Implicit is the assumption that true model is of fairly low dimension. Here, the criteria are derived to provide a consistent estimator of the order or dimension (K) of this "*true model*" and the probability

of selecting this "*true model*" approaches 1 as the sample size increases. Bozdogan (1987) provides an insightful review of many of the "*dimension consistent*" criteria. The best known of the "*dimension consistent*" criteria was derived by Schwarz (1978) in Bayesian context and is termed as BIC (Bayesian Information Criterion):

$$BIC = -2 \log L(\hat{\theta}) + K \log(n). \tag{1.8}$$

In the context of binary or censored responses, there is a debate about what n should be in the BIC approximation to the Bayes factor: the number of subjects or number of event? However, Volinsky (1997) provided evidence that n should be the total number of uncensored cases (deaths or events).

BIC arises from a Bayesian viewpoint with equal priors on each model and very non-informative priors on the parameters, given the model. The purpose of the BIC is to select the optimal model for simple prediction. BIC is not an estimator of relative K-L. It provides a consistent estimate of the true order of the model at the expense of assuming that a true model exists and is low-dimensional (Rissanen, 1989; Sclove, 1987).

1.4.6 Other Criteria

Rissanen (1989) proposed a criterion, called minimum description length (MDL), based on coding theory which is another branch of information theory. Hannan and Quinn (1979) derived a criterion (HQ) for model selection where the penalty term was $a(n) = c \log[\log(n)]$, n being the sample size and c a constant to be greater than 2 (Bozdogan, 1987). Their objective was to provide a consistent criterion in which

$a(n)$ is an increasing function of n at a rate as slow as possible. But this criterion has very little use in practice. Bozdogan (1987) proposed yet another criterion called $CAICF$ (C for consistent and F for Fishers information matrix):

$$CAICF = -2\log L(\hat{\theta}) + K\left\{\log(n) + 2\right\} + \log|I(\hat{\theta})|, \tag{1.9}$$

where $\log|I(\hat{\theta})|$ is the natural logarithm of the determinant of the estimated Fishers Information Matrix.

1.5 Summary and Discussion

In summary, for statistical inference, it is recommended to employ the class of information-theoretic criteria that are estimates of relative K-L information such as AIC or AICc for general use in the selection of a parsimonious approximating model (Burnham *et al.*, 1998). If we want to avoid over-fitting, then we should use the consistent criteria $cAIC$ and $CAICF$, sometimes at the cost of considerable under-fitting bias in smaller samples. Of course, as the number of observations gets large, for the consistent criteria, the probability of under-fitting on over-fitting will diminish. This suggests that one should use these consistent criteria for large samples. When the sample size is not very large, the dimension consistent criteria tend to select under-fitted models, which may result in considerable bias, over-estimated precision and associated problems in inference. Therefore, if we want to avoid under-fitting a model, then we should use AIC. If count data are found to be over-dispersed, then $QAIC$ and $QAICc$ are useful. For large samples, TIC

might offer an improvement over AIC or AIC_c. TIC is an asymptotically unbiased estimate of relative K-L information.

Many model selection criteria have been developed in recent years. Though hypothesis testing is widely used for model selection, there are problems in developing a general approach based on these tests. Relatively speaking, the concepts and practical use of the *information-theoretic* approach to model selection are simpler than those of statistical hypothesis testing, and much simpler than some of the Bayesian approaches to data analysis (e.g., Laud and Ibrahim, 1995; Carlin and Chib, 1995). It is well known that the classical theory of hypothesis testing focuses on determining whether observations from a given sample are consistent with some stated hypothesis or not. Thus, in the hypothesis testing tradition, frequently ignoring power considerations, we choose an arbitrary significance level α, for example the celebrated 5%, 2.5%, *or* 1%, and then try to determine (at least approximately) a critical value, from the standard tables of the test procedures, to make our decision. In these testing procedures there is no provision to penalize for over-parameterization since usually an unstructured saturated model is used as a reference (Akaike, 1987).

By contrast, when we use AIC, $cAIC$, or $CAICF$, the "*level of significance*" is adjusted in such a way that the corresponding probability of rejection of a simpler model decreases as degrees of freedom or complexity increases. This connection between model selection criteria and the level of significance α, provides us with a way to test the validity of different restrictions of a model. Also, it gives us

a yardstick in comparing every possible model and choosing the model giving the smallest probability of incorrect rejection of null hypothesis (i.e. *Type I error*) to be the best fitting model. This fact justifies the comparison of the model selection criteria in a class of models which are not nested and thus cannot be compared by the classical likelihood ratio tests. In my thesis, I will adopt this idea to select a parsimonious model and explore the behavior of *Type I error* while choosing the model *a posteriori*.

In the next Chapter, I will discuss in detail the issues related to model selection uncertainty which arises due to model building process.

Chapter 2

Model Selection Uncertainty

2.1 Introduction

"Model selection uncertainty" refers to the inference problems that arise when the same data are used for both model selection and inference related to this the *a posteriori* selected model. It is argued that model selection uncertainty should be fully incorporated into statistical inference whenever estimation is sensitive to model choice and that choice is based on the data used to estimate the model (Buckland *et al.*, 1997). If model selection uncertainty is ignored, precision is often over-estimated, achieved confidence interval coverage is below the nominal level, *type I error* rates in hypothesis testing are inflated, and predictions are less accurate than expected. The understanding that the model selection process might have a significant effect on the analysis leads to the acceptance of model selection uncertainty as an important aspect of data analysis.

Model uncertainty generally depends on two main inherent error sources: the model formulation error and input data errors. These types of errors may be es-

pecially important in the case of statistical modelling since there is no *type I error* assumptions imposed on a model structure. For every data set, a variety of potentially applicable model forms may exist and there is no universal model type to apply. In every case, a choice shall be made taking into account specific needs and available resources. It is suggested that the statistical model structure should be selected in the iterative way, when the created model is tested among possible alternatives, as proposed by Harvey (1981).

It should always be kept in mind that there is often considerable uncertainty in the selection of a particular model as the "*best*" approximating model. In realistic situations, there exist a best approximating model and other models that are poorer approximations due only to the finite amount of data available. Thus, one of the major source of uncertainty is the limited size of the sample. When we seek to fit models to our data, we shall begin to see how this uncertainty occurs. We should consider the consequences of limitations on the amount and quality of data available. At the center of this consideration is the idea of *sampling variability*. The observed data are conceptualized as random variables; their values would be different if another independent sample was available. It is this *sampling variability* that results in uncertain statistical inference from the particular data set. However, the effect of sampling variation is not only to reduce the precision with which we can fit the models, but also to limit the precision with which the "*optimal*" model may be identified. In fact, the existence of *sampling variability* means that the future

samples will vary from past samples and hence the model selected using the past samples may be different from the model that could be selected in a new sample. *sampling variability* thus causes uncertainty at all stages of modelling (Gilchrist, 1984).

Though literature on model selection methods has increased substantially in the past $15 - 20$ years, relatively little appears in the literature concerning the properties of parameter estimators in situations where a data-dependent model selection procedure has been used (Hurvich and Tsai, 1990; Goutis and Casella, 1995). Thus, little is known about the impact of using the same data to both select a proper model and to estimate the model parameters and their precision. Gilchrist (1984), Breiman (1992) and Chatfield (1995) gave insights into the problems when the same data are used both to select the model and to make inferences from the model.

If a best approximating model has been selected from a reasonable parsimonious set of candidate models, bias in the model parameter estimation might be small. However, there is uncertainty about which model to use. The model selection uncertainty is the component of variance in the estimation that reflects that model selection merely estimates the best approximating model, based on the single data set. A different model (in the fixed set of models considered) may be selected as the best for a different data set. For example, sampling variance of an estimator $\hat{\theta}$ given a model is usually estimated as $\mathrm{Var}(\hat{\theta}|\mathrm{model})$. However, if the model was selected from the same data, the actual sampling variance of $\hat{\theta}$ has two components:

1) Var($\hat{\theta}$|model) and 2) a variance component due to not knowing the *a priori* best approximating model (and therefore having to estimate this). Generally speaking, the variance will increase, as might be expected, from the additional uncertainty due to the model selection process.

If one uses a method such as AIC or cross-validation to select a parsimonious model given the data, and then estimates a conditional sampling variance given the selected model, the estimated precision will then likely be over-estimated, because the variance component due to model-selection uncertainty has been omitted. The standard errors (S.E.) computed conditionally on the model will be too small, confidence interval will be too narrow and achieved coverage will be below the nominal level (Chatfield, 1995; Rencher and Pun, 1980; Hurvich and Tsai, 1990; Potscher, 1991; Goutis and Casella, 1995; Kabaila, 1995).

The problem with model selection uncertainty shows a certain analogy to using the multiple coefficient of determination R^2 in multiple linear regression. Though R^2 can be estimated using standard least square (LS) regression theory based on a sample of n independent observations, it would not be the same with another sample of n independent observations. In fact, R^2 for a model is expected to be lower in an independent sample than in the sample from which the model was estimated. This is because the regression coefficients tend to be tailored for a particular data set.

On the other hand, the increasing complexity of empirical questions investigated through multi-variable modelling makes it less and less likely that the user will be

able to specify the model *a priori*. This implies that applied research will have to rely increasingly on models that are selected *a posteriori*. Therefore, the search for optimal methods for coping with model selection uncertainty is one of the priorities of current research on statistical inference. In the following sections, I shall review the effects of model uncertainty such as narrower confidence interval and biases in parameter estimates due to data-based modelling, and will also discuss the ways of assessing and overcoming the effects of model uncertainty through simulation studies and model averaging.

2.2 Confidence Subset of Models

Confidence subset of models can be defined to aid in identifying a subset of good models. The interpretation of a confidence interval of size $(1-\alpha)$ for a parameter of interest is very clear; i.e. in repeated samples from the process, $100(1-\alpha)\%$ of the data sets will generate a confidence interval that includes the true parameter value. This idea extends to that of generating a confidence subset of the models considered such that, with high relative frequency over samples, the subset of models contains the actual Kullback-Liebler (K-L) best model (defined below) among the entire set of models considered. The point is to make this subset as small as possible (analogous to narrow confidence intervals).

By K-L best model we mean the model with the shortest K-L distance between the true model f and candidate models g. To define K-L distance, let us suppose x denote the data being modelled and θ denote the parameters in the approximating

model g. Also let the function $g(x|\theta)$ be the approximating model g for data x given the parameters θ. Then the K-L distance is defined as (for continuous function) the integral

$$I(f,g) = \int f(x) \log \left\{ \frac{f(x)}{g(x|\theta)} \right\} dx.$$

K-L (1951) developed this quantity from "Information Theory". Here, $I(f,g)$ can be termed as the "information" lost when model g is used to approximate the reality (i.e. "*truth* f"). Hence according to this measure, one may seek an approximating model that loses as little "information" as possible; which is equivalent to minimizing $I(f,g)$ over the candidate models in the set. An equivalent interpretation to minimizing $I(f,g)$ is that we seek an approximating model that is the "*shortest distance*" away from truth.

Similar to confidence interval for a parameter based on a model and data, there exists the concept of a confidence set on the actual K-L best model. Several approaches were proposed to estimate the confidence subset of models based on information theory. I will discuss them in the following sections.

2.2.1 Method Based on Akaike Weight (w_i)

Akaike weights are the estimates of the relative likelihood of each fitted model in the set of models considered. After estimating the likelihood for each model (M_i, $i = 1, 2, ..., R$, where R is the number of models considered) given the data x, these likelihoods are normalized as follows to obtain the Akaike weight.

Specifically, the critical value for a confidence set of plausible models in terms of AIC difference (Δ_i) can be determined by extending the concept of the likelihood of the parameters given both the data x and model M_i ($i = 1, 2, ..., R$), i.e. $L(\hat{\theta}|x, M_i)$, to a concept of the likelihood of the model given the data , i.e. $L(M_i|x)$. Therefore, from Buckland *et al.*, (1997) we can write, $L(M_i|x) = C\, L(\hat{\theta}|x, M_i)$, where C is any arbitrary constant, $L(\theta|x, M_i)$ is the likelihood function of the parameter θ given the data x and the model M_i. Moreover, from the definitions,

$$\begin{aligned} \Delta_i &= AIC_i - \min AIC \\ &= -2\log L(\hat{\theta}|x, M_i) + C. \end{aligned}$$

We have therefore,

$$\begin{aligned} L(\hat{\theta}|x, M_i) &\propto \exp[-\frac{1}{2}\Delta_i], \quad \text{hence} \\ L(M_i|x) &\propto \exp[-\frac{1}{2}\Delta_i]. \end{aligned} \tag{2.1}$$

Now, to compare any model i with model j, the usual practice is to compute the relative likelihood as $\frac{L(M_i|x)}{L(M_j|x)}$. For a better interpretation of the relative likelihood, Burnham and Anderson (1998) suggested to normalize the $L(M_i|x)$ so as to obtain a set of positive "Akaike weights" w_i such that $\sum_{i=1}^{R} w_i = 1$. Thus the Akaike weights, commonly denoted as w_i, becomes,

$$w_i = \frac{\exp[-\frac{1}{2}\Delta_i]}{\sum_{r=1}^{R} \exp[-\frac{1}{2}\Delta_r]}. \tag{2.2}$$

Therefore the idea is, given that there are R models with one of them being the best, it is convenient to restrict the relative likelihoods to sum to 1. Obviously, for the

estimated K-L best model (say M_k), we will have $\Delta_k = 0$ and $\exp[-\frac{1}{2}\Delta_k] = 1$. So the odds for the ith model actually being the K-L best model is $\exp[-\frac{1}{2}\Delta_i]$ which can be defined as the "relative likelihood". By defining the Akaike weights (w_i) in this way, it is ensured that two models with the same value for AIC are given the same weight, whether or not they have the same penalty (number of parameters). This idea of the likelihood of the model given the data and the resulting model weights has been suggested in several articles mostly by Akaike (Akaike, 1978b, 1979,1980, 1981b, 1983b; Bozdogan, 1987; Kishino *et al.*, 1991).

Now, for the construction of the confidence set of models a rational (but not unique) approach is to sum these Akaike weights (w_i) from largest to smallest until that sum is just greater than or equal to $(1 - \alpha)$, α being the level of significance. Then the corresponding subset of models is a type of confidence set on the K-L best model (Burnham and Anderson, 1998). In this approach, the Akaike weights (w_i) are being interpreted as posterior probabilities (i.e. given data and the set of prior models) that model i is the K-L best model.

Interpretation of $\mathbf{w_i}$ Through an Example

From the equation (2.2), it is clear that, the bigger the value of Δ_i is, the smaller the w_i is and the less likely that model i is the actual K-L best model as illustrated on a hypothetical example in Table 2.1. It can be seen that for the 7 hypothetical models considered (ranked in order of increasing AIC i.e. decreasing goodness of fit), the normalized Akaike weights (normalized w_i) and the relative likelihoods ($\exp[-\frac{1}{2}\Delta_i]$)

Table 2.1: Relationship between AIC differences (Δ_i), Relative Likelihood $\exp[-\frac{1}{2}\Delta_i]$ and Normalized Akaike weights for 7 hypothetical models

Model No.	Δ_i	$\exp[-\frac{1}{2}\Delta_i]$	Normalized w_i
1	0	1.00000	0.43166
2	1.5	0.47237	0.20390
3	1.7	0.42741	0.18450
4	3.1	0.21225	0.09162
5	4.3	0.11648	0.05028
6	5.6	0.06081	0.02625
7	7.2	0.02732	0.01179

are inversely proportional to the AIC differences (Δ_i). I have normalized the Akaike weights (w_i) to show this proportional relationship. In this example, the best model among the candidate model is the model number 1 according to the rank provided by the normalized Akaike weight. However, note that the selected best model is only about twice as likely as each of the two next models. Thus, there will be a lot of variation in the selected best model from sample to sample. By contrast, e.g., models 6 and 7 have very low Akaike weights and are rather unlikely to be selected as the "best" models in any sample.

Bayesian Approach and Akaike Weights

In general, Akaike weights provide a better measure of the model plausibility than sampling theory based relative frequencies of model selection (Burnham and Anderson, 1998). Again, there is a Bayesian basis for interpreting the Akaike weight w_i as being the probability that model M_i is the K-L best model, given the data. Given

$L(M_i|x)$, i.e. the likelihood of the model M_i in a given sample, we can compute the posterior probability that model M_i is the K-L best model, provided that a specific prior probability distribution of the alternative models has been specified. That is, we must first specify a prior probability distribution $\tau_1, \tau_2, ..., \tau_R$ in which τ_i reflects our belief that model M_i will be the best K-L model. These probabilities τ_i must be specified independently of results of fitting any models to the data. So, τ_i is the one's prior degree of expected belief that model M_i is the true model or the degree of expected correctness of the model (Newman, 1997). Given a specific prior probability distribution, generalized Akaike weight (Kishino *et al.*, 1991) becomes

$$w_i = \frac{L(M_i|x)\tau_i}{\sum_{r=1}^{R} L(M_r|x)\tau_r}. \tag{2.3}$$

There may be occasions when unequal prior probabilities can be specified, based on substantive knowledge, justifying the use of (2.3) rather than its special case (2.2) in which all $\tau_i = \frac{1}{R}$ implying a non-informative prior.

However, the computation of weights w_i for the K-L best fitted model M_i does not represent a true Bayesian approach. A Bayesian approach to model selection requires both the prior τ_i on the model, and a prior probability distribution on the parameters θ in model M_i for each model. Then, the derivation of posterior results requires complex integration (usually only achievable by Monte Carlo methods) (Raftery *et al.*, 1993; Madigan and Raftery, 1994; Carlin and Chib, 1995). In fact, prior probability in (2.3), under information-theoretic approach to model selection, is not exactly the same as that of the Bayesian approach to model selection. The

Bayesian approach generally assumes that one of the models in the set of R models is true, or assumes that the weighted average of these R models corresponds to the true model, if model averaging is to be used. In contrast information-theoretic approach assumes that true model f is in the set of models and that $\tau_1, \tau_2, \tau_3, ..., \tau_R$ is a probability distribution of our prior information on the K-L best model for the data. In most applications, we believe that the issue can't be which model structure is true, because it is likely that none of the models considered is exactly true (Kapur and Kesavan, 1992). Rather, the issue is which model, when fitted to the data (i.e. when θ is estimated), is the best model for the purpose of representing the finite information in the data.

2.2.2 Method Based on Cut-off Δ_i

Based on the sampling distribution of AIC difference (Δ_i), $i = 1, 2, ..., R$, there is another way to develop the confidence set of models (Burnham and Anderson, 1998). Here the model producing minimum AIC is expected to be the K-L best model in the set of R candidate models. Roughly speaking, in the same spirit as $\hat{\theta}$ is the MLE of θ, the model corresponding to the biggest Δ_i can be considered as the *best* model. In a sampling theory context, Δ_i can be considered big if it is at or beyond the 95th percentile of the sampling distribution of a statistic analogous to $(\theta - \hat{\theta})$. To illustrate this idea, let index k corresponds to the actual expected K-L best model in the set, then $\Delta_k = AIC_k - \min AIC$, where Δ_k is a conceptual pivotal value whose sampling distribution is independent of any unknown parameters. This observable

random variable Δ_k is in fact, analogous to (θ-$\hat{\theta}$), which can often be used, after normalizing, as a pivotal value for the construction of a confidence interval on θ (Burnham and Anderson, 1998).

The sampling distribution of Δ_k has substantial stability and an alternative rule of thumb for an approximate 95% confidence set on the K-L best model is the subset of all models M_i having Δ_i less than or equal to some value, roughly between 4 to 7 (Burnham and Anderson, 1998). Such a subset of the models considered would include the actual K-L best model in 95% of all samples. Thus, according to Burnham and Anderson (1998), as long as observations are independent and sample sizes are large for any model i with $\Delta_i \leq 2$, there is no credible evidence that the model i should be ruled out as being the K-L best model for the population of all possible samples. For a model with $2 < \Delta_i \leq 4$, there is weak evidence that the model is not the K-L best model. If a model has $4 < \Delta_i \leq 7$, there is strong evidence that the model is not the K-L best model, and if $7 < \Delta_i \leq 10$, there is definite evidence that the model is not the K-L best model. Finally, in any situations if a model has $\Delta_i > 10$, there is strong evidence that this model is not competitive as the K-L best model (Burnham and Anderson, 1998). Thus, models with $\Delta_i > 10$ would not be generally included in the confidence set.

2.2.3 Method Based on Relative Likelihood

A third reasonable basis for defining a confidence set of models is motivated by the likelihood based inference (Edwards, 1992; Royall, 1997), which is analogous to

profile likelihood interval on a parameter given a model. Here the confidence set of models is all models i ($i = 1, 2, ..., R$) for which the ratio $\frac{L(M_i|x)}{L(M_k|x)}$ is "*small*" (such as $\frac{1}{8}$, Royall, 1997), where $L(M_i|x)$ is same as defined before and k is the index of the best model in the set. This criterion translates exactly into being the same as the set of all models for which Δ_i is less than or equal to some fixed cut-off point. For example, the cut-off value of Δ_i for which $L(M_i|x) \propto exp[-\frac{1}{2}\Delta_i]$ is small, is 0.135 (for $\Delta_i = 4$), 0.082 (for $\Delta_i = 5$) or 0.050 (for $\Delta_i = 6$).

2.2.4 Summary

Thus, there are three approaches to find a confidence set of models. The first is based directly on Akaike weights (w_i), summing as the probability of each model being the actual best model given the data (Burnham and Anderson, 1998). The second approach uses a cut-off Δ_i motivated by the idea of the sampling distribution of the approximate pivotal Δ_k (using the 95th percentile of this distribution as the cut-off). The third approach relies on relative likelihood. The first two methods are more popular and the third one, based purely on relative likelihood, is rarely used (Berger and Wolpert, 1984; Edwards, 1992; Azzalini, 1996; Royall, 1987).

2.3 Methods of Assessing Model Selection Uncertainty

It is likely that in selecting a parsimonious model for a given problem there will be substantial amount of sample to sample variation in the performance of alter-

native models on a given model selection criterion such as AIC or BIC, implying substantial model-selection uncertainty. Statistical inference should take this uncertainty at the model selection stage into account and should not be just based conditionally on the selected best model. Part of this inference process involves ranking the fitted models from best to worst and then going a step further for calibrating the relative plausibility of each fitted model (M_i) by a weight of evidence (w_i) relative to the selected best model (Burnham *et al.*, 1995). Then, using the full set of models and associated information such as conditional sampling variances and model weights, we can produce unconditional inferences over the entire set of models, such as unconditional sampling variances or model-averaged parameter estimates. Thus, model-selection uncertainty is a methodological problem in it's own right, well beyond just the issue of what is the best model. Therefore, instead of ignoring the uncertainty encountered while choosing a model, a better and logical approach might incorporate this uncertainty in subsequent inferences.

There are three general approaches to assess the model selection uncertainty; *i*) Monte-Carlo simulation studies *ii*) the bootstrap applied to a given set of data and *iii*) utilizing the set of AIC differences (Δ_i) from the set of models fitted to the data. Monte-Carlo and bootstrap are computer intensive. Whereas Monte-Carlo implies generating a large amount of simulated data sets, methods based on bootstrap and Δ_i values use a single data set.

The fundamental idea of the model based sampling theory approach to statis-

tical inference is that the data arise as a sample from some conceptual probability distribution f and hence uncertainties of our inferences can be measured if we can estimate f. There are different ways to construct a non-parametric estimator of f from the sample data. The Bootstrap technique as well as AIC can efficiently allow insight into model uncertainty.

2.3.1 General Comments on Bootstrap

The bootstrap is a type of Monte Carlo method applied case by case and based on realized data (Mooney and Duval, 1993). The most fundamental idea of the bootstrap method is that we compute measures of our inference uncertainty from the estimated sampling distribution of f. However, in practice bootstrap means using some form of re-sampling with replacement from the actual data x to generate B bootstrap samples x^*. Often the data consists of n independent units and then suffices to take a simple random sample of size n with replacement from the n units of data to get one bootstrap sample. Each sample gives an estimate of the unknown population parameter. The average of these values is called the bootstrap estimator and their variance is called the bootstrap variance. Thus, the set of B bootstrap samples is a proxy for a set of B independent real samples from f. From the set of results obtained in B bootstrap samples, we measure our inference uncertainties regarding the population. The bootstrap can work well for large sample size (n) but may not be reliable for small samples (e.g., $n \leq 20$) regardless of how many bootstrap samples are used (Burnham and Anderson, 1998).

2.3.2 AIC Differences (Δ_i), Model Selection Probability, and the Bootstrap

Using the bootstrap method we can estimate the sampling distribution of model-selection frequencies and the distribution of AIC difference (Δ_i). In this method, the role of the actual (unknown) K-L best model is played by the model selected as the best from the data analysis. Let M_k be the best model, on average, under the AIC selection criterion. For each bootstrap sample, we fit each of the R models, compute the corresponding AIC^*'s, and then find the single $\Delta_k^* = AIC_k^* - \min AIC^*$ ($*$ refers to the bootstrap sample). Here the $\min AIC$ and the value of AIC_k vary by sample; but k doesn't change over the bootstrap replications. For example, model M_6 might be the actual best model to always use ($k = 6$). Thus AIC_k^* is always the AIC value, from the given bootstrap sample for model k, which is the selected AIC best model for the data. The model producing $\min AIC^*$ varies by bootstrap sample. However, it is often the model k, that is the best model in a bootstrap sample, thus ${\Delta^*}_k = 0$ otherwise, ${\Delta^*}_k > 0$ i.e. when model k doesn't produce $\min AIC^*$.

The B bootstrap replications provide B values of ${\Delta^*}_k$ that are independent, conditional on the data. The percentile of the empirical probability distribution function of ${\Delta^*}_k$ across B bootstrap samples provide the estimate of the percentile of the sampling distribution of Δ_k and hence provide a basis for a confidence set on the K-L best model for the actual data.

For a $(1 - \alpha)100\%$ confidence set on the K-L best model, first we order the

$\Delta^*{}_k$ (smallest to largest) value to find $\Delta^*{}_{k.(b)}$ for $b = [(1-\alpha)B]$. For the actual data analysis results, the subset of the R models M_i having $\Delta_i \leq \Delta^*{}_{[(1-\alpha)B]}$ is the desired confidence set. The number B of bootstrap replication needs to be $10,000$ or higher for reliable results (Burnham and Anderson, 1998). Other information can be gained from these bootstrap results about model selection uncertainty, in particular, the frequency of selection of each of the R models. Thus, the estimator of the relative frequency of model selection in the given situation is $\hat{\pi}_i = \frac{b_i}{B}$, where b_i is the number of replications in which model i is selected as the K-L best model. These estimated selection probabilities are useful for assessing how much sampling variation there is in the selection of the best model; they directly quantify model selection uncertainty. These estimated selection probabilities are similar, but not identical in meaning, to the Akaike weights (w_i), which also quantify strength of evidence about model selection uncertainty.

For each bootstrap replication we can compute the Akaike weights w_i^* ($*$ refers to bootstrap replication) and then average these over the B replications to get $\overline{w}^*{}_i$ (Burnham and Anderson, 1998). Comparison of w_i, $\overline{w}^*{}_i$ and $\hat{\pi}_i$ is informative, each of which provides information about the sampling uncertainty in model selection.

In fact, the theoretical measure of model-selection sampling uncertainty is the set of true unknown selection probabilities $\pi_1, \pi_2, \ldots\ldots, \pi_R$. Either $\hat{\pi}_i$ (from the bootstrap) or the Akaike weights w_i may provide a basis to estimate the uncertainty about model selection, given a single sample.

2.3.3 Monte Carlo

The Monte Carlo method is a universal numerical method of approximately solving mathematical and physical problems by the simulation of random quantities and/or by random sampling (Sobol, 1994). One attractive feature of this method is the simple structure of the computation algorithm. As a rule, a program is written to carry out one random trial, and is then repeated N times, each trial being independent of the others. Finally the results of all trials are averaged. In this method, the model generating the data is assumed to be included in the set of candidate models (Burnham and Anderson, 1998). Practically, in Monte Carlo investigations, $10,000$ to $100,000$ independent samples are generated from the "*true*" model. Then, one applies model selection to each sample and summarizes resulting relative frequencies of models selected and other information of interest, such as variation of the Δ_i and unconditional variances of parameter estimators. The important difference between bootstrap and Monte Carlo is that while both approaches imply direct observation of the sample to sample variation in model selection, Monte Carlo technique requires specifying the "*true*" data-generating model, whereas bootstrap relies entirely on the sample data.

2.4 Uncertainty in Parameter Estimates Associated With Model Selection

Parameter estimation uncertainty is conceptually separable from model selection uncertainty. Given a correct model an MLE is reliable and we can compute a

reliable estimate of its sampling variance and a reliable confidence interval (Royall, 1997). If the model is selected entirely independently of data at hand and is a good approximating model, and if n is large, then the estimator of the sampling variance is essentially unbiased and any appropriate confidence interval achieves its nominal coverage. This would be the case if we use only one model selected on *a priori* basis, and this model g fits well the data generated under the true model f.

However, in the case of data-based model selection, the selection process is expected to introduce an added component of sampling uncertainty into any estimated parameters, which leads to too small classical estimates of sampling variances. These classical estimates are conditional on the model and do not reflect model selection uncertainty. Confidence intervals based on such a conditional model can be expected to have coverage rate lower than nominal coverage (Hurvich and Tsai, 1990; Abrahamowicz *et al.*, 1996).

Consider a scalar parameter θ which may be used in all or only in some of the models considered, but is in the selected model and therein has unknown value θ_i given model M_i. Here, the subscript i denotes the model used to estimate θ with the understanding that this parameter means the same thing in all models in which it appears. Consider a true value of θ, which would be estimated from the true model f, even though θ need not to literally appear in f. Given model M_i, the MLE $\hat{\theta}_i$ has a conditional sampling distribution, and hence a conditional sampling variance $\mathrm{Var}(\hat{\theta}_i|M_i)$.

The ideas of classical sampling theory can be used to derive the theoretical sampling variance of $\hat{\theta}$ resulting from the two-stage process of (i) model selection, and (ii) estimating $\hat{\theta} \equiv \hat{\theta}_i$ given that model M_i was selected (Buckland *et al.*, 1997). Repeating this process m times on independent samples produces a direct estimate of the unconditional sampling variance, unconditional meaning the variance estimate does not dependent on one model being absolutely correct. Thus, the estimated unconditional variance becomes

$$\widehat{Var}(\hat{\theta}) \equiv \sum_{j=1}^{m} (\hat{\theta}_j - \bar{\hat{\theta}})^2/(m-1), \tag{2.4}$$

where $\bar{\hat{\theta}}$ is the simple average of all m estimates. This variance estimator represents the total variation in the set of m values of $\hat{\theta}$; hence both within- and between-model variation is included. This set of m values of $\hat{\theta}$ can be partitioned into R subsets, one for each model wherein the ith subset contains all the $\hat{\theta}$'s computed from the samples in which the model M_i was selected. Then, we can compute, from the ith subset of the $\hat{\theta}$ value, an estimate of the conditional sampling variance of $\hat{\theta}$ when model M_i was selected. Burnham and Anderson (1998) gave the formula for estimating Var($\hat{\theta}$) as a weighted combination of conditional variances, plus a term of variation among θ_1,.....,θ_R. The weights involved are in fact the model selection probabilities. Relevant formulas are shown in Section 2.4.1. If a given parameter appears only in some of the models, the basis for unconditional inference about that parameter can be based on just the relevant models (Buckland *et al.*, 1997). An example is variables selection in multiple linear regression of a dependent variable

y on p regressors, x_1, x_2,...,x_p. There are 2^p possible models, but each regressors appears only in a half of these models (i.e. 2^{p-1} models). Thus, if parameter β_j, corresponding to regressor variable x_j, is in the selected best AIC model, we must restrict ourselves to just the subset of models that contain β_j in order to directly estimate the unconditional sampling variance of $\hat{\beta}_j$ (Burnham and Anderson, 1998).

2.4.1 Including Model Selection Uncertainty in Estimating Sampling Variance

The variance component due to model selection uncertainty should be incorporated into estimates of precision to obtain correct unconditional standard errors of regression parameters. These estimators of unconditional variance are also appropriate in cases where one wants a model-averaged estimate of the parameter of interest. Assume that θ appears in all models considered; then we define a model-averaged parameter value θ_a as $\theta_a = \sum_{i=1}^{R} \pi_i \theta_i$, where π_i is the probability of selecting model M_i in repeated sampling. So the estimator of θ is $\hat{\theta}_a = \sum_{i=1}^{R} \hat{\pi}_i \hat{\theta}_i$. Here θ_a is not necessarily the same as θ, which could be estimated only if the *true* model f was known. Under classical sampling theory, $\hat{\theta}$ arrived at the two stage process of model selection and the parameter estimation given the model, is by definition an unbiased estimator of $\hat{\theta}_a$. Therefore, the unconditional sampling variance of $\hat{\theta}$ has to be measured with respect to θ_a. The theoretical unconditional sampling variance of the estimator of θ is given by

$$Var(\hat{\theta}) = \sum_{i=1}^{R} \pi_i \left[Var(\hat{\theta}_i | M_i) + (\theta_i - \theta_a)^2 \right]. \tag{2.5}$$

The quantity $Var(\hat{\theta}_i|M_i) + (\theta_i - \theta_a)^2$ is just the mean square error (MSE) of $\hat{\theta}_i$ given model i. Thus, in one sense, the unconditional variance of $\hat{\theta}$ is just an average MSE across the R models (Burnham and Anderson, 1998). Specifically,

$$E[(\hat{\theta}_i - \theta_a)^2|M_i] = Var(\hat{\theta}_i|M_i) + (\theta_i - \theta_a)^2.$$

This quantity can be considered as the sampling variance of $\hat{\theta}_i$ given model M_i, when $\hat{\theta}_i$ is being used as an estimator of θ_a. By averaging across models, we get formula (2.5) which incorporates model selection uncertainty into the estimated variance of $\hat{\theta}$. Accordingly, to get an estimator of $Var(\hat{\theta})$, estimated values could be plugged in (2.5) to get

$$\widehat{Var}(\hat{\theta}) = \sum_{i=1}^{R} \hat{\pi}_i \left[\widehat{Var}(\hat{\theta}_i|M_i) + (\hat{\theta}_i - \hat{\theta}_a)^2\right]. \tag{2.6}$$

Ignoring the fact that π_i and $Var(\hat{\theta}_i|M_i)$ are estimated, we can evaluate $E[\widehat{Var}(\hat{\theta})]$ to bias-correct $\widehat{Var}(\hat{\theta})$ (Burnham and Anderson, 1998). The result involves the sampling variance of the model averaged estimator $Var(\hat{\theta}_a)$, and is given by

$$E[\widehat{Var}(\hat{\theta})] = Var(\hat{\theta}) + \sum_{i=1}^{R} \pi_i Var(\hat{\theta}_i|M_i) - Var(\hat{\theta}_a),$$

Hence,

$$Var(\hat{\theta}) = Var(\hat{\theta}_a) + \sum_{i=1}^{R} \pi_i E(\hat{\theta}_i - \hat{\theta}_a)^2. \tag{2.7}$$

From (2.7) it is clear that the use of $\widehat{Var}(\hat{\theta}) = \widehat{Var}(\hat{\theta}_a) + \sum_{i=1}^{R} \pi_i (\hat{\theta}_i - \hat{\theta}_a)^2$ may considerably inflate the variance compared to $\widehat{Var}(\hat{\theta}) = \widehat{Var}(\hat{\theta}_a)$ (Burnham and Anderson, 1998).

However, if our goal is to estimate θ_a, then the model averaged $\hat{\theta}_a$ is to be preferred to $\hat{\theta}_i$ because it will have a smaller sampling variance.

From Buckland *et al.*, (1997), we can write the unconditional sampling variance of θ_a as

$$Var(\hat{\theta}_a) = \left[\sum_{i=1}^{R} \pi_i \sqrt{Var(\hat{\theta}_i|M_i) + (\theta_i - \theta_a)^2}\right]^2, \tag{2.8}$$

with the corresponding estimator as

$$\widehat{Var}(\hat{\theta}_a) = \left[\sum_{i=1}^{R} \hat{\pi}_i \sqrt{\widehat{Var}(\hat{\theta}_i|M_i) + (\hat{\theta}_i - \hat{\theta}_a)^2}\right]^2. \tag{2.9}$$

Thus, we can say that the Akaike weights (w_i) or bootstrap probabilities ($\hat{\pi}_i$) that are used to rank and calibrate models can also be used to estimate unconditional precision where interest is in the parameter θ over R models. In terms of Akaike weights (w_i) the estimated variance of $\hat{\theta}$ becomes

$$\widehat{Var}(\hat{\theta}) = \left[\sum_{i=1}^{R} \hat{w}_i \sqrt{\widehat{Var}(\hat{\theta}_i|M_i) + (\hat{\theta}_i - \hat{\theta}_a)^2}\right]^2. \tag{2.10}$$

These estimators include a term for the conditional sampling variance $\widehat{Var}(\hat{\theta}_i|M_i)$ and incorporate a variance component for model selection uncertainty $(\hat{\theta}_i - \hat{\theta}_a)^2$. The standard practice of the conditioning on a single selected model ignores model uncertainty which leads to the under-estimation of uncertainty in subsequent inferences. A Bayesian approach to this problem involves averaging over all possible models while making inferences about quantities of interest. Thus, in order to incorporate model selection uncertainty into inference, we can consider the philosophy of model averaging by weighting the alternative models, rather than selecting between

them. Going one step further, we assume that the fitted models are in some sense a random sample from an infinite set of possible models, each of which provides a valid estimate of the parameter in its own right.

2.5 Variable Selection in Multiple Regression and Model Uncertainty

Variable selection is often the main focus of model selection in the context of multiple regression models. It has been acknowledged by many authors (e.g., Kleinbaum *et al.*, 1988; Neter *et al.*, 1990) that different subset selection strategies can indicate different "*best*" models. This is due to several factors, including the order in which the variables are entered into the model and the criterion used to evaluate the models. This situation clearly indicates the existence of model uncertainty. In fact, choosing a single model from the set of models indicated by a variable selection procedure and making inferences as if this model was the true model disregards model uncertainty.

One important issue to be cared for here is to quantify the evidence for the importance of each variable, or each relevant subset of variables. As for example, let us consider that we have 10 models corresponding to different combinations of a number of possible regressor variables. Now we assume that the selected best model includes x_1 and has an Akaike weight of only 0.3 (see Equation 2.2 in Section 2.2.1). There is a lot of model-selection uncertainty here, and hence there would seem to be only a weak evidence for the importance of variable x_1 based on the selected best

model. But to quantify the importance of x_1 we must consider the Akaike weights (w_i) of all other models that include x_1. It might be that all models that exclude x_1 have very low Akaike weights; which would suggest that x_1 is a very important predictor here. To measure this importance one has to sum the Akaike weights (or the bootstrap probability $\hat{\pi}_i$) for all relevant models.

Although model selection is widely recognized as central to good inference, paradoxically it has seldom been integrated fully into inference (Buckland *et al.*, 1997). For example, there are many methods in multiple regression for identifying an appropriate subset of covariates. Once covariates are selected, subsequent inference is usually conditional on the selected model. The reason that inference is generally conditional on the selected model is the complexity of unconditional inference.

It is well known that model selection stage can severely affect the validity of standard regression procedures. Rencher and Pun (1980) demonstrated that a model, selected by the best subset regression method, tends to have an inflated value of R^2. Breiman (1988) showed that models selected by various data-driven methods can produce strongly biased estimates of mean squared prediction error. Reviews of some of the difficulties induced by variable selection were given by Bancroft and Han (1977) and Miller (1984). The latter author showed that if one starts with a model selected from the data, then regression estimators may be biased and standard hypothesis tests may not be valid.

Hurvich and Tsai (1990) discussed, as follows, the impact of model selection on

Table 2.2: Coverage Rates for Confidence Intervals, Conditional on Selected Model with Nominal Rate $(1-\alpha)$, $n = 20$, Correct Model, $p_0 = 3$

Model No. (p)	$(1-\alpha) = 0.90$	$(1-\alpha) = 0.95$	$(1-\alpha) = 0.99$
3	0.901	0.955	0.987
4	0.672	0.836	0.918
5	0.717	0.804	0.913
6	0.622	0.865	0.946
7	0.568	0.727	0.885
OCR[a]	0.806	0.900	0.960
Z − score[b]	-7.006	-5.130	-6.742
ICR[c]	0.910	0.952	0.973

[a] Overall Coverage Rate
[b] Z-score is based on normal approximations to binomial distributions under null hypothesis of nominal coverage rate
[c] Initial Coverage Rate
Source : Hurvich and Tsai (1990), Table 1, pp 215.

inference in multiple linear regression. A common practice is that once a model has been selected, one analyses the data as if they were a fresh data set. Thus, conditionally on the event of having selected a particular model, the distribution of the data may be substantially different from their unconditional distribution. The authors presented Monte Carlo evidence that such a difference does indeed exist and explored the impact of this difference on the coverage rates of confidence intervals for the regression parameters. Specifically, they performed 500 realizations generated from 3 normal linear regression models

$$Y = X_0\beta_0 + \epsilon, \tag{2.11}$$

where Y is an $n \times 1$ vector of observation, X_0 is an $n \times p_0$ matrix of explana-

tory variables, β_0 is a p_0 dimensional parameter vector, and ϵ is an $n \times 1$ vector of *i.i.d.* standard normal random variable. The three models generated were $n = (20, 30, 50)$, $p_0 = (3, 4, 4)$, $\beta_0 = [(1, 2, 3)'; (1, 2, 6)'; (1, 2, 3, 6)']$. In each case, 7 candidate variables were stored in an $n \times 7$ matrix X of *i.i.d* standard normal random variables. The candidate model of dimension p consisted of the first p columns of X. For each realization, AIC and BIC (see Chapter 1 for details) criterion were employed to select a value of p (dimension or order of models). The study was focused mainly on the three interpretations of coverage rates, namely nominal, conditional and overall interpretation. The authors found that conditional coverage rates were much smaller than the nominal coverage rates when the model was known in advance. Some of these results are reproduced in Table 2.2 which shows the conditional coverage rates for nominal 90% 95% and 99% confidence region (based on percentage point of F distribution) with the models selected by AIC. It is clear from the results of Table 2.2 that when the correct model is selected ($p = p_0$), the conditional coverage rates are close to nominal rates, while for over-fitted model ($p > p_0$), the conditional rates are substantially smaller than the nominal rates (results not shown here). So the obvious impact is that conditional rate tends to be substantially below the nominal rates when the model is over-fitted. It was also shown that under-fitting ($p < p_0$) results in zero conditional coverage rates (not shown here). Given these results, the authors suggested that model selection and inference should be performed on separate parts of data (i.e. "data splitting") and argued that this

approach would ensure same conditional and nominal coverage rates.

2.6 Model Selection In Survival Analysis

Despite their importance for biostatistics, only a small proportion of the model selection research has focused on survival regression models. The censoring, typical of survival data, adds another level of complexity, possibly compounding the potential bias in failing to account for model uncertainty. Some authors (Elston and Johnson, 1994; Forthofer, 1995) stated that stepwise methods are often used for variable selection in survival regression models. Kalbfleisch and Prentice (1980) emphasized for careful attention while choosing variables in the model, and for its importance to account for prognostic factors in the model even when there is no significant statistical evidence of their connections to the response in the data at hand. Flemming and Harrington (1991) employed a standard variable selection process in their classic example of the analysis of prognostic factors in primary billiary cirrhosis (PBC) dataset. First, they calculated a Rao test statistic for each variable individually. Next, a step down procedure eliminates 5 of the 11 original variables. A likelihood ratio test verified that the removal of those five variables was not significant ($p = 0.2$). The logarithms of the remaining variables were then included and another stepwise procedure led to the transformation of 3 of the remaining 5 variables. The final model contained 5 variables, including the three transformed ones. However, in all these applications of model selection in survival analysis, the variance of the regression parameters are under-estimated by not accounting for the

uncertainty in the model selection procedure.

Altman and Anderson (1989) described a bootstrap experiment of stepwise methods for Cox's model. They performed a stepwise selection to a on PBNC dataset (different from Flemming and Harrington) and identified 6 of the 17 independent variables as significant. Applying stepwise to bootstrap samples of the data, they tested the "*stability*" of the model selection process. One hundred applications of stepwise on bootstrap samples of the data resulted in widely different sets of significant variables. Each of the 17 variables was significant in at least one of the bootstrap runs, while 4 of the 6 "*significant*" variables from the full analysis were found to be significant in fewer than 75% of the bootstrap runs. This could be interpreted as an evidence that all of the variables are important, but some have stronger evidence from the data than the others. It seems unwise to throw away variables because they do not meet arbitrary "significance level". These results demonstrate the instability of variable selection in Cox's model. The analyses of Kuk (1984) and Raftery *et al.* (1995) also showed that model uncertainty plays a large role in Cox's model.

However, the problems related to *a posteriori* model selection have to be balanced against the potential benefits. In a study, Abrahamowicz *et al.*, (1996) used regression splines (Ramsay, 1988; Wegman and Wright, 1983) to model the hazard ratio as a flexible function of time to overcome the restrictive proportional hazards assumption. To determine the flexibility of the spline estimate, they used both

a priori fixed model (model with highest degrees of freedom) and the minimum AIC model selection approaches. It was clear that minimum AIC criterion tended to select more parsimonious models, reduced over-fitting bias, and stabilized the estimates. On the other hand, the coverage rates of the AIC-optimal models were found to be well below the nominal level, thus reflecting the inability of conventional inference to account for additional variance due to *a posteriori* model selection.

2.7 Summary

Model selection should not be considered just as the search for the best model. Rather the basic idea ought to be to make more reliable inferences based on the entire set of models that are considered *a priori*. This would imply ranking and calibrating the set of models and possibly determining a confidence subset of models for the K-L best model. Parameter estimation should use all the models within the confidence subset by averaging over models. Finally, unconditional variances should be used to quantify the uncertainty about parameters of interest unless the selected best model is strongly supported.

In general, there is a substantial amount of model selection uncertainty in many practical problems. Such uncertainty about model structure and associated parameter values, arises to the K-L best model, whether one uses hypothesis testing, information theoretic criterion, dimension-criteria, cross validation, or various Bayesian methods (Burnham and Anderson, 1998). Usually, the uncertainty reported for values such as future predictors or parameter estimates consists only of the uncer-

tainty associated with the statistical distributions embedded in the model. Ignoring the uncertainty associated with the model selection procedure results in an under-estimation of all uncertainty and leads to over-confidence in reported conclusions (Volinsky, 1997).

Model selection uncertainty can be quantified in two basic ways: it can be based on the differences in AIC values for the set of models considered, or it can be directly based on the bootstrap methods. If there is substantial model selection uncertainty and if the sampling variance is estimated conditionally on the selected model, the actual precision of estimated parameter will likely be over-estimated and the achieved confidence interval coverage will be below the nominal level (e.g. Abrahamowicz *et al.*, 1996; Hurvich and Tsai, 1990). Estimation of unconditional variances can be made using either Akaike weights (w_i) or bootstrap selection probabilities. The bootstrap provides direct, robust estimates of the model selection probabilities π_i. Thus, when there may be suitable analytical or numerical estimations of conditional (on a given model) sampling variances, the bootstrap may be used to get unconditional measures of precision. Otherwise, Akaike weights will help estimating unconditional sampling variances. In any case, a carefully thought out set of *a priori* models should eliminate model redundancy problems and is a central part of a sound strategy for obtaining reliable inferences (Burnham and Anderson, 1998).

The importance of identifying a small number (R) of candidate models defined

prior to detailed analysis of the data cannot be overstated. In the case of all possible combinations of possible regressors, if one has p candidate variables, then $R = 2^p$ so that with large p the number of candidate models can exceed the size of the data set.

Finally, investigators should explain what procedure was actually employed in the model selection. Was it based on objective model-selection inference applied to *a priori* identified set of candidate models? Alternatively, was the selected model a result of a subjective strategy of seeking a model that fits well the data by introducing new models into consideration as data analysis progresses? In the former case, either AIC or its variants is recommended, as needed. In the latter case, if the strategy can be implemented in a computer algorithm, then the use of bootstrapping is suggested to asses the model selection uncertainty (Burnham and Anderson, 1998).

In Chapter 3, I will discuss the model selection problem in the specific area of selecting the "*optimal*" functional form of the dose-response relationship in Cox's proportional hazards model (1972).

Chapter 3

Model Selection Problems in the Context of Choosing the Optimal Transformation of a Continuous Covariate in Cox's Regression

3.1 Introduction

Prognostic models are tools which are intended to predict the average course of a disease given the values of covariates known as prognostic factors. Prognostic factors may be represented by binary, categorical or continuous covariates. Though continuous variables are common in all fields of application, it is often unclear how to handle them as explanatory variables in regression models. The researcher faces with the challenge of building a reliable regression model, must decide how to deal with the continuous factors. In the following sections, I discuss various issues related to the use and modelling of continuous covariates in regression analysis along with the simulation study I carried out for my thesis.

3.1.1 Representation of the Effect of a Continuous Covariate as a Specific Area Where Data-Dependent Model Selection is often Employed

In epidemiology, various strategies can be used for analyzing the effect of a quantitatively measured (continuous) exposure variable, on the risk of developing a certain disease. Examples of exposure variables commonly measured on a continuous scale include the number of cigarettes smoked per day or during a lifetime, alcohol intake, ionizing radiation, dietary energy intake, and arsenic concentration in water. A common problem in the statistical analysis of clinical studies is the selection of these variables in the framework of a regression model, which might influence the outcome variable. Investigations of the stability of a selected model are often called for, but usually are not carried out in a systematic way. Since analytical approaches are extremely difficult, data-dependent methods might be an useful alternative (Sauerbrei and Schumacher, 1992).

A large part of clinical research consists of studies of prognostic factors, which identify covariates that are related to survival or can predict disease outcome. It is often the case that potential prognostic variables are continuous in nature and the functional relationship of the covariate with survival is explored, to evaluate how the risk of death varies, as the value of the covariate changes. Identification of a group of independent prognostic variables may be done through various modelling techniques, such as stepwise procedures. Once identified as a prognostic variable, a continuous covariate may need to be *dichotomized* or *categorized*, which is often done by using an

"*optimal*" cut points i.e. cut points that maximize risk difference, or other statistic related to model's fit, for a given sample (Schulgen *et al.*, 1994). But this method may over-estimate the prognostic importance of the variable resulting in invalidating the obtained *p values*. So, there will be a considerably higher risk of *detecting* a significant effect of a variable that is in reality not prognostic, resulting in an inflated *type I error* rate (see for example Halpern, 1982). Indeed, data-dependent decisions related to the choice of the "*optimal*" categorization of a quantitative variable are likely to induce "*optimistic*" bias. However, the absence of a simple method that gives a good representation of different forms of curvilinear relationships may be one reason for the common practice, especially in medical statistics, of converting continuous variables into ordinal variables with two or more categories (Royston and Altman, 1994).

In fact, common practice of handling continuous prognostic variable in clinical and epidemiological studies is still limited to very simplistic methods. Inclusion of simple linear terms and categorization are currently the most widely used strategies to deal with continuous covariates in multiple regression models (Brenner and Blettner, 1997). Although during the past decade, alternative strategies, such as polynomial regression or spline regression (Ramsay, 1988; Ramsay and Abrahamowicz, 1989; Maclure and Greenland, 1992; Greenland, 1995) for dealing with continuous covariates in multiple regression models have received increased attention, they are seldom used in practice. Existing alternatives such as cubic splines and non-

parametric smoothers have a large potential but have also important drawbacks; they are computationally intensive, they do not yield compact expressions for prediction, thus can't be fitted by using standard regression software and there may be difficulty in explaining them to the non-expert users (Royston and Altman, 1994).

Fractional polynomials provide another methodology that can extract important prognostic information which the traditional approaches may miss. Although fractional polynomials have been used by various researchers on an *ad hoc* basis (e.g., Isaacs *et al.*, 1983; Guo *et al.*, 1988) and recently by Sauerbrei and Royston (1999), only few references were found in the standard text books on regression. For example, Snedecor and Cochran (1967) mentioned the addition of 'terms like $\sqrt{Z}$, $\log(Z)$ or $\frac{1}{Z}$, if the data had required it', but they didn't elaborate further. Similarly, Draper and Smith (1981) discussed briefly *reciprocal*, *logarithmic* and *square-root* transformations of covariate, but did not pursue the topic. Atkinson (1985) described how constructed variables may be used to detect the need of covariates' transformation but did not consider fractional polynomials as such.

From the above overview, it is clear that the selection of the continuous covariate is related to the standard problem of model selection in regression analysis. In general, the interpretation of results from any exposure-response analysis depends on the choice of both the exposure functions and the model (Vacek, 1997). In the following section, I will discuss the usual approaches to modelling the effects of continuous covariates in epidemiology.

3.1.2 Overview of Conventional Approaches to Modelling the Effects of Continuous Covariates in Epidemiology

A challenge in epidemiological studies is to determine the form of the relationship between a given continuous risk factor and the risks of a disease. Various approaches have been described for analyzing continuously measured exposure in this context. Exposure can be modelled continuously or it can be categorized. In conventional general linear models, the effect of continuous covariates are typically assessed assuming linearity of true dose-response curve, possibly after adding an approximate link such as logit for binary outcomes (McCullagh and Nelder, 1989). Identification of a linear dose-response relationship can add support to a causal association based on classic criterion of causation (Hill, 1965). Not all associations, however, are linear in nature. One way to avoid the linearity assumption, often adopted in epidemiological studies is to categorize continuous variables. Categorization has the advantage that it is easy to interpret and is more robust with regard to outliers and model misspecification, but it has the possible disadvantage of loss of efficiency (Breslow and Day, 1980, 1987; Zhao and Kolonel, 1992). However, owing to imprecise measurement and lack of prior knowledge about the functional shape of the relation between exposure and risk of disease, the strategy used most commonly for analyzing quantitative exposure seems to be categorization or even dichotomization of exposure (Schulgen *et al.*, 1994). The analysis of the resulting categorical variables implies assumption of a constant effect within the defined categories of exposure

(Brenner and Blettner, 1997).

Another approach is to assume that the effect of exposure has a special functional shape on the (log) odds ratio or (log) relative risk of developing the disease; exposure then can be analyzed as a continuous covariate in a regression model. While the latter approach avoids the definition of cut-points, the assumption of a linear or quadratic effect, or any other pre-specified functional shape, over the whole range of exposures may be questionable (Schulgen *et al.*, 1994). Categorization of continuous covariates has been proposed to allow for more flexible modelling of the shape of covariate-risk association in such situations (Rothman, 1986).

However, as noted by others (Brown *et al.*, 1994; Weinberg, 1995), there are limitations too, when using categorized exposure. As an alternative, parametric models like polynomial or more flexible fractional polynomial and spline regression models has been suggested (Royston and Altman, 1994). Though in recent years fractional polynomials came to the forefront of dose-response analysis, it has been argued that spline regression has the advantage over more traditional linear regression methods, in that it may be regarded as an approximation to non-parametric regression and therefore is relatively insensitive to the subjective choices of modelling parameters such as number of pieces or order of spline (Greenland, 1995). On the other hand, non-parametric models (e.g. generalized additive models) may well fit the data but can be difficult to interpret due to fluctuations in the fitted curve. In conclusion, while both the methods of fractional polynomials and of the spline regression can

be valuable when important non-linearities are anticipated, they are rarely used in studies of health effects of alcohol, serum cholesterol, nutrients and other prognostic factors.

3.1.3 Overview of Conventional Regression Models For Predicted-Response Relationships

Because of the inability to experimentally control for the differences between values of important factors observed in individual study participants, confounding is a major concern in most epidemiological and clinical studies, except the randomized controlled trials (Rothman, 1996). For this reason, multiple regression models are commonly applied in clinical and epidemiological studies, to prevent from confounding bias. Two models have come to the forefront as they accommodate the types of responses that commonly occur in clinical/epidemiological studies: binary responses such as in-hospital death or presence/absence of a certain condition, and censored continuous responses such as the time until death or until a therapeutic response. Such data are typically analyzed using logistic regression models and Cox's proportional hazards model (1972), respectively. These models provide powerful analytic tools that yield valid statistical inferences and make reliable predictions if various underlying assumptions are satisfied. Two types of assumptions underlying a variety of regression models concern the distribution of the response variable and the nature or shape of the relationship between the predictors and the response (Harrell *et al.*, 1988).

In several types of modelling situations, for example linear regression and logistic regression, we may find transformations of the continuous independent variables (or covariate) desirable. The multiple logistic regression model belongs to the broad family of parametric generalized linear models that rely on the assumption that the effects of continuous predictors, possibly after applying some conventional transformation, are linear (linearity assumption) (McCullagh and Nelder, 1989). The linearity assumption simplifies both model estimation and interpretation of results, since it allows for summarizing the effects of a continuous predictor by a single parameter (e.g., in the logistic model, the logarithm of odds ratio corresponds to a one unit increase in the predictor value). However, the linearity assumption may be too simple to represent the effect of some risk factors correctly. In many practical applications, there is no *a priori* justification for this assumption. In the case of departure from linearity, for a given risk factor, parametric logistic regression estimate will under-estimate its effect over some range of values and over-estimate the effect over some other range (Abrahamowicz *et al.*, 1997). To address these issues, various flexible non-parametric statistical methods for estimating the true shape of the regression function or for assessing whether a postulated shape is correct are employed.

One of the main challenges in the non-parametric regression is to determine the "*optimal*" degree of flexibility, which also determines the bias/variance trade-off (Abrahamowicz *et al.*, 1996; Ramsay, 1988). In fact, two common approaches are

possible. First, the model may be fixed *a priori*, to ensure the flexibility sufficient to approximate all functions of potential interest (Gray, 1992; Ramsay, 1988). Second, *a posteriori* model selection criterion such as *AIC* can be used to find a reasonable trade off between model parsimony and fit to the data at hand and to reduce the risk of over-fitting bias (Abrahamowicz *et al.*, 1992; Sleeper and Harrington, 1990). In this thesis, I use the latter approach that is to rely on the minimum *AIC* to select *a posteriori* the "*optimal*" transformation of a continuous predictor in proportional hazards model, and then estimate the impact of this approach on *type I error* through simulations.

Before presenting the simulations, I briefly discuss the logistic regression and Cox's proportional hazards regression model in the next two sub-sections.

3.1.4 Logistic Regression Model

Logistic regression is a flexible statistical modelling approach for binary responses that permits many types of inferences, including treatment comparisons and prediction. It has a major advantage over older methods such as discriminant analysis in that it allows for a direct estimation of probability with no assumptions about distributions of the variables (Harrell *et al.*, 1988). A brief description of the model follows.

Let $Z = (Z_1, Z_2, ..., Z_p)'$ be a vector of predictor or dependent variables. For a binary response variable Y with values 0 or 1, the logistic regression model (Cox, 1958) is stated in terms of probability that the event $Y = 1$ occurs given the de-

scriptor value of $z' = (z_1, z_2, ..., z_p)$:

$$\text{Prob}\{Y = 1|Z = z\} = \frac{1}{1 + \exp\{-(\beta_0 + \beta' z)\}}, \tag{3.1}$$

where, β_0 is an intercept and $\beta' = (\beta_1, \beta_2, ..., \beta_p)$ is the vector of regression coefficients corresponding to the descriptors.

The logistic model can be restated as a linear model by using the logit of $\text{Prob}(Y = 1|Z = z)$, where logit denotes $\log[\frac{P}{(1-P)}]$:

$$\text{logit}\,[\text{Prob}\{Y = 1|Z = z\}] = \beta_0 + \beta' z. \tag{3.2}$$

When the relationship is not linear, one can employ transformations of the covariate to satisfy the linearity assumption (Hosmer and Lemeshow, 1989)

Other versions of logistic model are available for ordinal or polytomous responses (Hosmer and Lemeshow, 1989).

3.1.5 Proportional Hazards Regression Model

There exist many ways of incorporating independent variables into a model that uses hazard rate as the dependent variable, including parametric models like exponential and Weibull models (Kalbfleisch and Prentice, 1980). Cox (1972) developed an important and widely used semi-parametric method called the proportional hazards (PH) model. Cox's model is a popular choice for the analysis of censored survival data because it is semi-parametric, i.e. avoids the need to specify the unknown distribution of the time to event (i.e. survival time), conceptually appealing, and efficient against PH alternatives. The fundamental assumption of Cox's model is

that hazards are proportional, i.e. their ratios remain constant over the entire follow-up time, where the constant is determined solely by their covariate vector. Another basic assumption of the PH model is that the conditional log hazard function is an additive function of time and of the vector of covariates. That is, the modelled response is the hazard rate of failure, with a log hazard ratio (HR) that is linear in the covariates. However, this assumption is violated when covariate effects are best represented by smooth, nonlinear functions. In Section 3.2, where I introduce the design of my simulation study, I describe a flexible survival model that does not require linearity of the covariate function.

Survival data typically arise in a clinical trial setting. Patients enter the trial at random times during the accrual phase of the study, and their times to failure are observed (Sleeper and Harrington, 1990). When the data are analyzed, the observation times of all patients who have not yet failed are considered censored. The data collected on each patient ($i = 1, 2, ..., n$) are of the form $(t_i, \delta_i, \mathbf{z_i})$, where t_i, δ_i and $\mathbf{z_i}$ are defined as follows. The datum t_i (survival time) equals $\min(T_i, C_i)$, where T_i and C_i are independent random variables denoting the "true failure time" and censoring time, respectively. The indicator δ_i takes value 1 if failure is observed (i.e. if $T_i \leq C_i$) and 0 otherwise, and the vector $\mathbf{z_i}$ contains covariates or prognostic factors ($Z_{1i}, Z_{2i}, ..., Z_{pi}$, p being the number of covariates) that are thought to affect survival.

The PH model introduced by Cox (1972) is typically written as:

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\exp(\beta'\mathbf{z}), \tag{3.3}$$

where $\lambda(t|\mathbf{z})$ is the conditional hazard rate of failure given $\mathbf{z}$ at time t (i.e. instantaneous risk of death), $\lambda_0(t)$ is the 'baseline' hazard (i.e. hazard function for a "standard" individual, when all elements in the covariate vector $\mathbf{z}$ equal zero), and β is the corresponding vector of regression coefficients. These coefficients can be estimated by partial likelihood approach, and the inverse of the observed Fishers information matrix provides an estimate of their variances and covariances. We can linearize the above model by dividing both sides of the Equation (3.3) by $\lambda_0(t)$ and then taking natural logarithm of both sides as

$$\log \frac{\lambda(t|\mathbf{z})}{\lambda_0(t)} = \beta'\mathbf{z}. \tag{3.4}$$

Here, $\hat{\beta}$ in the fitted PH model is the estimated change in the log of the HR when the value of $\mathbf{z}$ is increased by 1 unit.

Cox's model can also be stated in terms of the survival function that describes the probability that the event will not occur before time t, thus

$$S(t|\mathbf{z}) = \mathrm{Pr[T > t|z]} = \mathrm{S_0(t)}^{\exp(\beta'\mathbf{z})}, \tag{3.5}$$

where $S_0(t)$ is the underlying baseline survival function for the "standard" individual.

The most general form of the PH model is

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\exp[g(\mathbf{z})], \tag{3.6}$$

where g is an unspecified function (Hastie and Tibshirani, 1986). Although this model does not restrict the log HR to be linear in $\mathbf{z}$, it is usually difficult to estimate $g(\mathbf{z})$ and to interpret the influence of any single covariate on survival. An additive regression model (Stone, 1985) provides more structure but allows a different, arbitrary function for each covariate. In the case of the Cox model, the log hazard ratio may have p components, each represented by an arbitrary function as:

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\exp[\sum_{j}^{p} g_j(z_j)] \tag{3.7}$$

where functions $g_j(z_j)$ may change their analytical form depending on the covariate $j = 1, 2, ..., p$.

In many instances a covariate transformation g can be approximated reasonably well by a polynomial. However, splines, i.e. piecewise polynomials, are well known for their success in interpolating and their usefulness in providing a smooth approximation to a covariate function of unspecified form (Wold, 1974). When a continuous covariate affects the log hazard in a smooth fashion, a spline function is a natural choice for approximating the covariate transformation (Sleeper and Harrington 1990). O'Sullivan (1988) and Gray (1992) used smoothing splines to estimate nonlinear covariate effects in the Cox model, while Sleeper and Harrington (1990) used regression splines. The generalized additive model (GAM) of Hastie and Tibshirani (1987) provides a powerful method for detecting nonlinear covariate effects in data. All regression models with response density belonging to the exponential family are in the GAM class. However, inference on GAM estimates is not well

developed, particularly in the case of the Cox's model (Sleeper and Harrington, 1990). Similar inferential problems occur in other spline-based models for flexible estimation of the covariate effects in the PH model (Kooperberg *et al.*, 1995).

Traditionally, in the PH model and in some other survival analysis models, the dependence of the survival time on covariates is modelled fully parametrically, so that the conditional hazard regression function can be estimated independently on the baseline hazard function (Cox and Oakes, 1984; Kalbfleisch and Prentice, 1980).

In summary, the above review indicates that there is an increasing interest in modelling nonlinear effects of continuous covariates in Cox's model. Such modelling can reveal new practically important aspects of the covariate effect on hazard (see e.g. Sleeper and Harrington, 1990; Gray, 1992; Abrahamowicz *et al.*, 1997). It can also reduce problems related to the arbitrary categorization of a continuous predictor (Schulgen *et al.*, 1994). However, flexible non-parametric methods based on smoothing or regression splines such as GAM (Hastie and Tibshirani, 1990) create difficult problems with the inference about the estimates (Kooperberg *et al.*, 1995; Abrahamowicz *et al.*, 1996). For this reason, a common practice is to estimate several alternative models, each using a different parametric transformations of the covariate of interest and then to select *a posteriori* the transformation that offers an *optimal fit* according to a criterion such as minimum AIC or maximum likelihood (Quantin *et al.*, 1999). However, this common approach can be considered a specific case of the general problem of data-dependent model selection and, therefore, is likely

to induce some bias at the step of statistical inference, as discussed in Chapter 2 of this thesis. In the next section, I investigate this issue through a simulation study.

3.2 Design of the Simulation Study

The problem considered here is to account for model selection uncertainty in the setting of Cox's model, with focus on the uncertainty involved in the selection of the "*optimal*" transformation of a continuous predictor. In this setting, my aim is to evaluate the impact of model selection on the *type I error* and *statistical power* through simulation studies.

In this section, I describe the design of the numerical study that I have carried out. I generated the survival data assuming two different situations. The first part of simulation, in which hazard was independent of the covariate of interest, focused on the *type I error* rate. The second part, where covariate was assumed to affect the hazard according to a pre-specified parametric function, investigated issues related to *statistical power*. The data generation procedure is described as follows.

3.2.1 Data Generation Procedure

Here, I describe in a general way, how I simulated the data sets conditionally on the chosen covariate values and/or censoring pattern. We considered the case of only one covariate. A covariate value z_i, a survival time T_i, an observed event time t_i, and a censoring time C_i are associated with each individual i ($i = 1, 2, .., n$). Let F, F_z, and F_c be the distributions of the survival time, the covariate, and the

censoring time, respectively. The survival data sets in our simulation studies were generated according to the following steps:

1. Get one set of data (t_i, δ_i, z_i), $i = 1, 2, ..., n$, as follows: First generate the covariate values z_i from F_z, then the "true" survival times T_i from F and the censoring times C_i from F_c. Next form the data points (t_i, δ_i, z_i), where $t_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, which equals 1 if the subject i is dead at t_i and 0 if it is censored.

2. Repeat step 1 m times, to form the m simulated samples.

For each combination of relevant parameters, I generated $m = 1000$ independent samples.

From the foregoing description of the simulated data set, it is clear that the following factors affect our simulation:

1. The covariate distribution F_z

2. The survival time distribution F

3. The regression parameter vector β, corresponding to log HR associated with unit a increase in the covariate

4. The form of the censoring distribution F_c

5. The average percent censoring

6. The sample size n

7. The particular dose-response function g, linear or non-linear, expressing the effect of covariate on the log *hazard*.

3.2.2 Considered Configurations

From the previous seven points following configurations were considered;

1. F_z uniform $(0, 1)$ distribution

2. F standard exponential distribution, with $\lambda_0(t) = 1$ for all t. This distribution is a special case of the Cox model (Equation 3.3).

3. $\beta = 0$ (HR=1) for the type I error and different parameter values for the power, which are resulting from the point seven below.

4. F_c exponential distribution

5. F_c with mean $(\frac{1}{\lambda})$, where $\lambda = 1.5$, resulting in approximately 37% average amount of censoring

6. $n = 100$ and $n = 300$

7. The different functional forms of the covariate considered are described below.

Form of The Dose-Response Curve

In the case of dependence of hazard on z, the hazard function can be expressed as

$$\log\ \lambda(t|z) = 1 + g(z), \tag{3.8}$$

where g is the function expressing the form of the dose-response curve. The function g includes the parameter of association between the covariate and survival time. Obviously, here the distribution of the time of death is an exponential with mean $\frac{1}{\lambda(t/z)}$.

I used the following 3 functions of z. The first function has a linear form.

$$g(z) = (z - v), \tag{3.9}$$

where v is used for scaling purpose only, and is fixed to 0.5. The following two functions have nonlinear form. The first function is a quadratic function:

$$g(z) = c(z - v)^2, \tag{3.10}$$

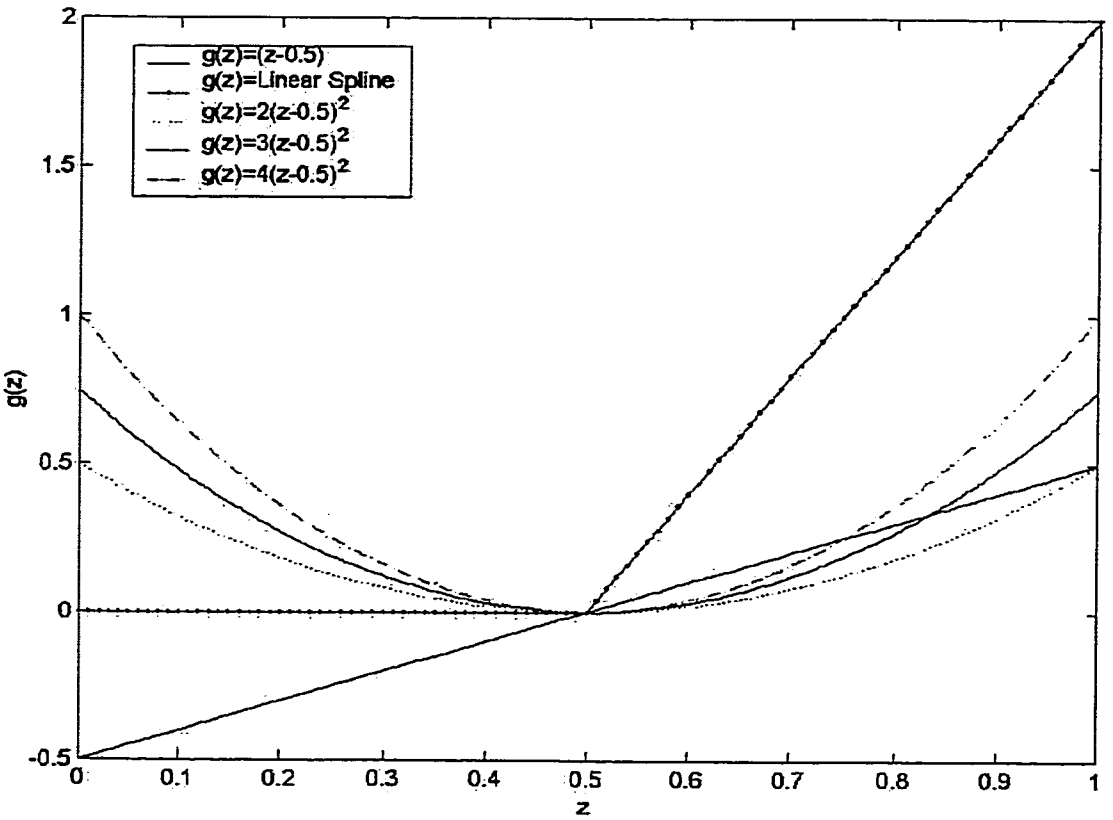
where v is fixed to 0.5 as above and c varies from $c = 2$ to $c = 4$, in order to investigate whether the results depend on the strength of the covariate effect, which influences the empirical *statistical power* of all tests. The second nonlinear function is a linear spline (bi-linear function):

$$g(z) = \begin{cases} 0 & \text{if } z \leq v \\ 4(z - v) & \text{otherwise} \end{cases} \tag{3.11}$$

where v represents a threshold intensity and again is fixed to 0.5, below which exposure has no effect. Here, fixing v at zero yields the corresponding function $g(z)$ to have no threshold, since covariate z is positive ($z \sim U(0, 1)$).

Figure 3.1 shows the five functions representing the alternative dose-response relationships considered in our simulations.

Figure 3.1: Functions Representing the Linear and Non Linear Dose-Response Relationship



There is a practical consideration behind choosing these particular functions. The linear dose-response function (Function 3.9) corresponds to a standard assumption underlying most of *Generalized Linear Models*. In the case of simulations, when this function is assumed to actually represent the "*true*" form of the association, it is interesting to investigate what is the impact of considering additional functional forms, in terms of *statistical power*. In other words, I will assess "*what price*", in terms of loss of power, the data analyst is expected to "*pay*" for considering additional models, which in this case are not expected to systematically improve fit to data.

By contrast, the simulations in which $g(z)$ assumes a non-linear form (Function 3.10 and 3.11) will help assessing to what extent different hypothesis testing procedures based on *a posteriori* selection of the "*optimal*" model may help detecting a statistically significant association.

The quadratic U-shaped function (Function 3.10) was selected to mimic the effects of those continuous covariates for which the risks are the lowest in the middle range, corresponding to "*typical*" or "*normal*" values, and increase in both tails of the distribution. This occurs often in clinical and epidemiological studies and a classic example is the effects of Body Mass Index (BMI) on cardiovascular risks (Abrahamowicz *et al.*, 1997; see also Chapter 4). Indeed, both obese individuals (high BMI) and those under-weighted (low BMI) have increased risks.

Finally, the broken line (linear spline) form (Function 3.11) is considered to

represent a frequent situation when increasing values of a risk factor does not affect the outcome in the low to middle range of its distribution, but beyond a certain threshold the risks start to increase sharply.

3.2.3 Details of the Data Generation Algorithm

In this section, I describe how I simulated the survival time and censoring time distributions, which are exponential. The details of the procedure is as follows.

Let us consider a continuous random variable Y with distribution function F_y. Also let U be a uniform$(0, 1)$ random variable. Now, according to the well known *inverse transformation* method (Ross, 1997), one can generate the random variable Y from the continuous distribution function F_y by generating a random number U, and then setting,

$$Y = F^{-1}(U) \tag{3.12}$$

Our assumption of a constant failure rate (i.e. hazard is constant throughout the time) implies an exponential density function of survival time (T). Therefore, our survival model is one parameter exponential distribution ($\exp(\lambda)$), from which we can generate the survival data by utilizing the relationship (3.12).

So, if $T \sim \exp(\lambda)$, then for any time point $t \geq 0$, the corresponding probability density function ($p.d.f.$) and cumulative distribution functions ($c.d.f.$) are,

$$f(t) = \lambda e^{-\lambda t}$$

and

$$F(t) = 1 - e^{-\lambda t},$$

respectively. Using (3.12), we can write

$$t = F^{-1}(u),$$

which leads to

$$u = F(t) = 1 - e^{-\lambda t}.$$

Hence,

$$\lambda t = -log(1 - u).$$

Since $(1 - u) \sim U(0, 1)$, I generated t using

$$t = -\frac{1}{\lambda} log(u). \tag{3.13}$$

Now, under exponential model, survival function associated with hazard $\lambda(t|\mathbf{z}) = \lambda$ can be written as

$$S(t|\lambda) = 1 - F(t|\lambda) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}.$$

So, the survival time (T) corresponding to $S(t|\lambda) = u$ can be determined (using 3.13) as the solution to the equation $S(t|\lambda) = u$ and hence

$$T = -\frac{1}{\lambda} log(u). \tag{3.14}$$

Under the assumption stated above, the survival time for an individual with covariate z is generated by substituting λ in Equation (3.14) by the hazard rate function conditional on the covariate

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\exp[g(\mathbf{z})], \tag{3.15}$$

where g is an arbitrary function of z, which may be linear or non-linear. Using similar arguments, censoring time C was generated as

$$C = -\frac{1}{\lambda_c}\log(u), \tag{3.16}$$

where λ_c is an arbitrary parameter of the censoring distribution (we considered 1.5). It should be emphasized that although both T_i and C_i are generated from an exponential distribution, these two values are completely independent of each other because the uniform variate u is generated each time separately. Thus, random right censoring, independent of z and T values, is introduced by generating censoring times C_i from an exponential distribution, whose parameter is set so as to obtain moderate censoring (here 37%).

The final survival data was constructed by comparing the corresponding individual T_i and C_i values: if $T_i \geq C_i$ then the individual is censored at time C_i, otherwise the individual dies at time T_i.

We generate all the random variables using S-Plus (1995) statistical package. The program code used to generate and to analyze the simulated data, is included in Appendix **A**.

3.2.4 Data Analytical Procedures

All the simulated samples were analyzed using the same general approach. First, seven different versions of Cox model, each corresponding to a different functional transformation of the covariate $g(z)$, were fitted to each generated sample. Since, there were no substantive grounds to select the "*correct*" functional form of the covariates $g(z)$ *a priori*, we considered the following seven simple parametric functions often used in epidemiological research: z, z^2, z^3, $\exp(z)$, $\log z$, $\sqrt{z}$, $\frac{1}{z}$. Then, under the PH assumption, we tested the hypothesis of no association between the covariate and hazard separately for each transformation, by using the likelihood ratio (LR) test, which asymptotically follows χ^2 distribution with 1 degree of freedom (d.f.). Conventional significance level of $\alpha = 0.05$ was used for all tests.

The main focus of the analysis was on testing the null hypothesis of no association between the covariate and hazard (HR=1), and on comparing the performance of various testing procedures. First, we compared the procedure based on one model at a time with a two-step procedure in which at the first step the minimum AIC model was selected and then the LR test was carried out using this "*best AIC*" model's results. Using the first approach, for each combination of relevant simulation parameters, we obtained the distribution of $1,000$ p-values of the LR test, each corresponding to one simulated sample, separately for each of the 7 functional forms of the covariate. In addition, the overall distribution of all $7,000$ models and sample-specific p-values was obtained by pooling the results obtained with 7 sepa-

rate models. In the second approach, the distribution of $1,000$ p-values was obtained using for each sample only the model that turned out to be the best fitting model for that sample, according to AIC criterion. It should be noted that the minimum AIC-based selection implied that for different samples the p-values represented, in fact, the models based on different functional forms of the covariate transformation. In addition to estimating the overall distribution of p-values, I also estimated the proportion of p-values smaller than the nominal significance level of $\alpha = 0.05$; together with the corresponding 95% confidence intervals. For simulations in which the covariate has no effect on hazard (HR=1), this proportion represents the actual size of the test, i.e. the empirical *type I error* rate generated by a given testing procedure. In simulations where the covariate does influence the hazard (HR $\neq 1$), this proportion allows us to estimate the *empirical power* of various testing procedures.

Preliminary Investigations of Some New Approximate Approaches to Hypothesis Testing Based on Model Averaging

In survival analysis, data-dependent model selection is commonly employed in analyzing the data and in making inference about the parameters of interest. In Chapter 2, I have reviewed the problems that data-dependent model selection techniques induce at the step of statistical inference and demonstrated that it is essential to account for model uncertainty. A suitable model averaging provides a framework to account for the model uncertainty and to improve estimation and inference. In fact, combining the results of many models allow the statisticians to take advantage of

the strength of different models and directly address the model uncertainty *inherent* in *a posteriori* selecting a single model. For instance, Breiman (1996) suggested to use a type of averaging to stabilize the inferences.

In view of this recommendation, I considered a simple approach to model averaging based on Akaike weights (w_i) that were discussed in Section 2.2.1. Specifically, I employed two versions of the model-averaged LR test. First, the un-weighted statistic was calculated as:

$$LR_{un-weighted} = \frac{1}{7}\sum_{j=1}^{7} LR_j \tag{3.17}$$

where LR_j indicates the value of the conventional LR statistic for the jth model. Next, the weighted statistic was calculated based on AIC weights:

$$LR_{weighted} = \frac{1}{7}\sum_{j=1}^{7} w_j LR_j, \tag{3.18}$$

where

$$w_j = \frac{\exp[-\frac{1}{2}\Delta_j]}{\sum_{r=1}^{R}\exp[-\frac{1}{2}\Delta_r]}.$$

Each of these two statistics (Equation 3.17 and 3.18) were calculated for each generated sample.

The theoretical distributions of the two resulting test statistics under the null hypothesis of no association are unknown. In this preliminary investigation, their empirical distributions, estimated from the results of simulations with no effect of the covariate, will be compared with the chi-square distribution with 1 degree-of-freedom. This comparison will help assessing to what extent such simple approaches

to model averaging may reduce problems related to the inflation of the *type I error* rate due to *a posteriori* model selection.

Chapter 4

Results

4.1 Results of the Simulations Study

Employing the procedure described in Chapter 3, I generated 1000 samples for different combinations of relevant parameters, including both the case of no association between covariates and hazard, and the case when the covariate does influence the hazard. Then the data from each simulated sample were analyzed using 7 different versions of Cox's model corresponding to the 7 functions of the covariate considered (see Section 3.2.4). The results are presented in the following sections.

4.1.1 Assessing the Impact of Model Selection on Type I Error Rate

Table 4.1 summarizes the results of simulations in which the covariate does not have any effect on the hazard ($HR = 1.0$), for the first of the two approaches described in section 3.3.4. Each of rows of Table 4.1 corresponds to a different model, which uses a particular parametric function to represent the covariate effect. For each model and two different sample sizes considered ($n = 100$, 300), first 4 columns of Table

4.1 show the observed type I error rate, corresponding to the LR test, at the nominal significance level of $\alpha = 0.05$.

The first two columns of Table 4.1 correspond to the first approach mentioned in section 3.2.4 (Unselected Model). So the results obtained are the *type I error* rates calculated from the distribution of $1,000$ p-values of each sample with 7 different functions of covariates considered. As expected, for each model the rates are quite similar to the nominal rate of 0.05. The last two rows show the overall *type I error* rates, pooled from all the $7,000$ estimates (7×1000 simulations) and the corresponding 95% confidence interval. The overall rate agrees well with the nominal test size as the 95% confidence interval includes 0.05 for both sample sizes.

The middle part of Table 4.1 shows the model-specific *type I error* rates (based on the 2nd approach mentioned in Section 3.2.4), calculated based on only those samples for which a given model was selected as the minimum AIC model. This simulates the situation in which, first the 7 different models are estimated for the same data, and then the minimum AIC model is used for testing the null hypothesis of no association. In a clear contrast to the left part of Table 4.1, in the case of data-dependent AIC-based model selection, the *type I error* rates for all models are much higher than the nominal 0.05 rate. In fact, the observed overall proportion of incorrect rejections across all models is about $0.14 - 0.15$, regardless of sample size, i.e. is almost 3 times higher than expected for $\alpha = 0.05$. The fact that the 95% confidence interval for the proportion of H_0 rejections begins at, or above, 0.10

Table 4.1: Observed Type I Error Rates of The LR Test For The Nominal Size $\alpha = 0.05$ and Proportion of Samples Where a Given Model Was Selected as the Best AIC Model

Model	Unselected Model Approach		Best AIC Model Approach			
	Observed Type I Error Rate		Observed Type I Error Rate		Proportion of Samples When the Model is Selected	
	n = 100	n = 300	n = 100	n = 300	n = 100	n = 300
Z	0.06	0.05	0.17	0.11	0.06	0.06
Z^2	0.05	0.06	0.17	0.18	0.05	0.07
Z^3	0.05	0.07	0.12	0.11	0.26	0.22
expZ	0.06	0.06	0.33	0.32	0.01	0.02
$\log Z$	0.05	0.05	0.13	0.18	0.14	0.18
$\frac{1}{Z}$	0.08	0.07	0.17	0.15	0.40	0.37
$\sqrt{Z}$	0.05	0.05	0.12	0.09	0.08	0.08
Overall	0.059	0.057	0.148	0.144	1.00	1.00
95% C.I.	0.044-0.074	0.043-0.071	0.126-0.170	0.100-0.180		

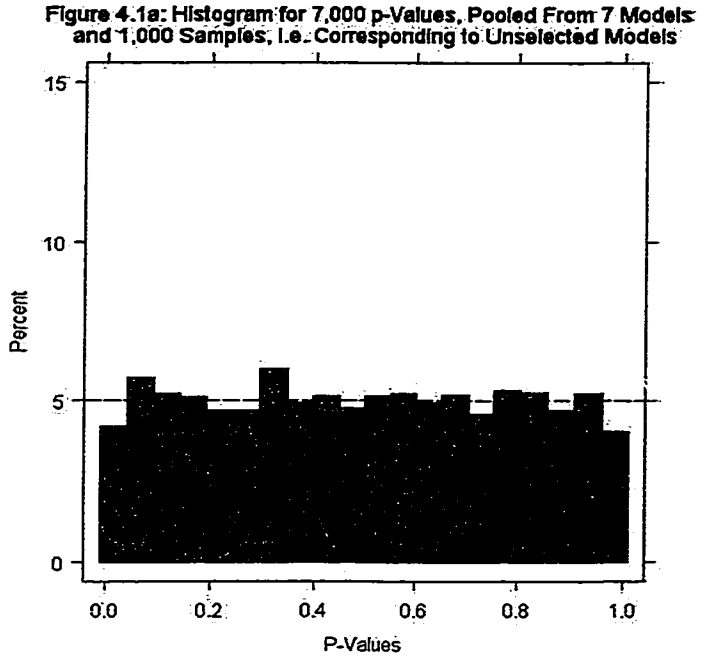
implies that the inflation of *type I error* rate due to AIC based model selection is statistically very significant.

The last two columns of Table 4.1 provide some insight into the problem. They show the proportion of times that each covariate function was selected as the best AIC model. For the 1000 samples, it is found that the function $\frac{1}{Z}$ occurred most frequently (40%), followed by Z^3 (26%) and $\log Z$ (14%) and changing the sample size (n) did not affect these proportions. The fact that each of the 7 models is selected as the minimum AIC in at least one sample indicates considerable model uncertainty (It should be noted that in the case of H_0 being true, this uncertainty is "*inherent*" as none of the model is better than any other). Thus, the model selected for inference about association between the covariate and hazard varies from sample to sample. More importantly, the model is not selected at random but so as to minimize the AIC criterion, i.e. to maximize the fit to data, in terms of the log likelihood. Therefore, LRT statistics are systematically higher than among un-selected samples and, as a consequence, the corresponding p-values are systematically lower than expected.

This systematic bias in the distribution of p-values for the sample size $n = 300$ is illustrated in Figure 4.1. The Figures show four distributions of p-values for the LR statistics in the simulations where the covariate has no effect on hazard (corresponding to results in Table 4.1). The Figure 4.1a shows that the distribution of all $7,000$ p-values, pooled from $1,000$ simulations and 7 models, is quite close to

the uniform distribution, which is expected given that the null hypothesis is true. By contrast, Figure 4.1*b* shows that the distribution of $1,000$ p-values corresponding to minimum AIC models in subsequent samples is substantially skewed to right, the low p-values are systematically over-represented while the high p-values are very rare. Overall, all the results clearly indicate that regardless of sample size, data dependent *a posteriori* model selection, is prone to considerable inflation of *type error* rates because of the failure to account for model selection uncertainty.

Figure 4.1: Figures Showing Systematic Bias in the Distribution of p-values



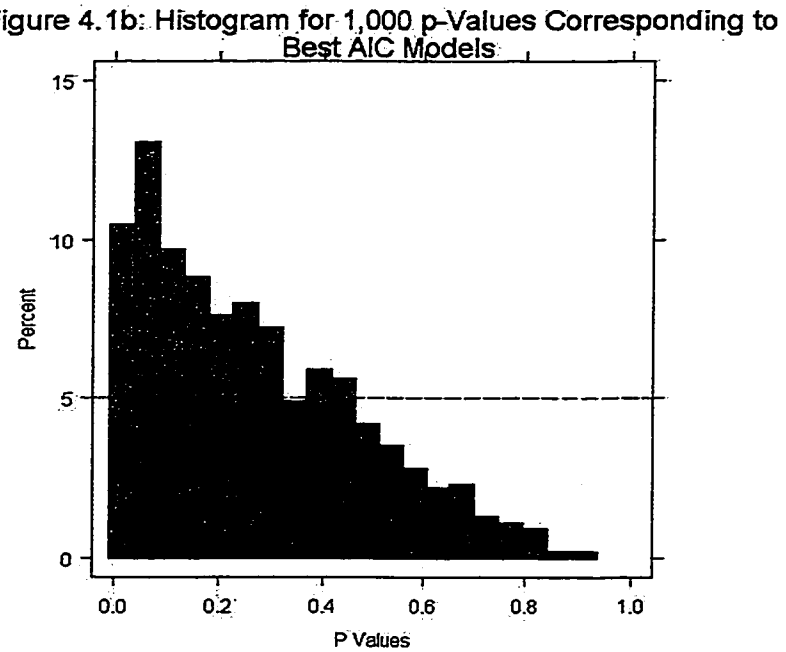


Figure 4.1c: Histogram for 1,000 p-Values Corresponding to Best AIC (Weighted) Models

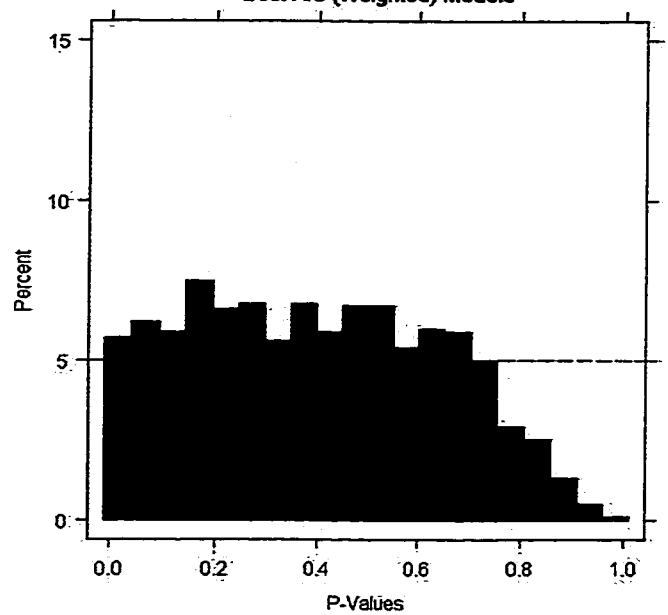
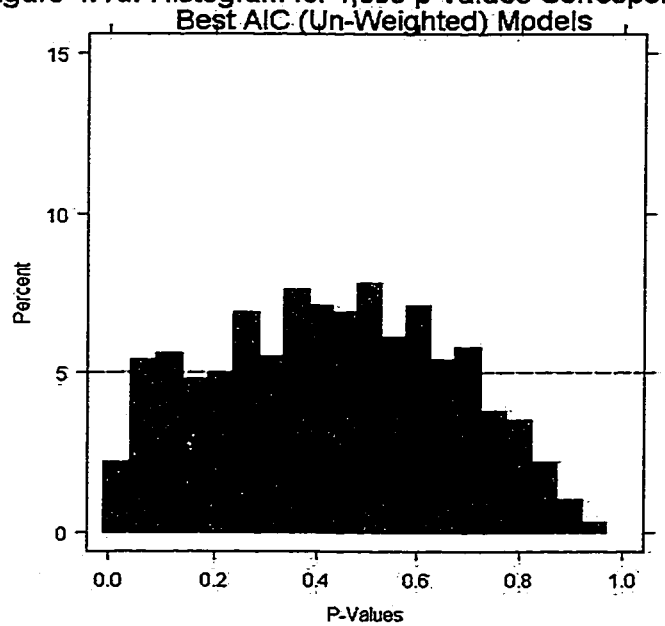


Figure 4.1d: Histogram for 1,000 p-Values Corresponding to Best AIC (Un-Weighted) Models



4.1.2 Preliminary Evaluation of Two Simple Model Averaging Approaches to Testing the Hypothesis of No Association

Given, on one hand, the serious inflation of type I error due to *a posteriori* model selection (Table 4.1 and Figure 4.1*b*) and, on the other hand, the practical difficulties of specifying the "correct" functional form of the dose-response curve, it is important to search for alternative approaches to hypothesis testing in the presence of model uncertainty. Therefore, in this section we evaluate two very simple heuristic approaches that are quite easy to implement and yet may be expected to reduce the magnitude of type I inflation (see Section 3.2.4 for the description of the two approaches). Table 4.2 compares the overall type I error rates, for the simulations with no covariate effect, obtained with four testing procedures. The two left-most parts of the Table 4.2 simply replicate the overall results from Table 4.1, based on un-selected testing and AIC-based model selections, respectively. The third part of Table 4.2 shows that the weighted version of model averaging based on Akaike weights (see Section 3.2.4) is able to reduce the observed type I error rate to half of the level obtained with the best AIC model-based testing (0.08 vs. 0.15). The rate yielded by the weighted approach is still too high, as indicated by the lower bound of the 95% confidence interval exceeding 0.05 but the overall *type I* level is much more acceptable than is the case of testing based on the minimum-AIC model. The right most part of Table 4.2 show that the un-weighted version of model averaging, in which the test statistic is constructed by simply taking the arithmetic

Table 4.2: Proportion of Simulated Samples, Where H_0 was Rejected at $\alpha = 0.05$, When There is No Covariate Effect: Effect of Model Averaging on Type I Error

n	Unselected Model	Best AIC Model	Model Averaging	
			Weighted	Un-weighted
100	0.059 $(0.044 - 0.074)^a$	0.148 (0.126-0.170)	0.08 (0.065-0.099)	0.03 (0.023-0.045)
300	0.057 (0.043-0.071)	0.144 (0.100-0.180)	0.08 (0.060-0.094)	0.03 (0.017-0.037)

[a] 95% Confidence Interval for the Proportion of p-values < 0.05

mean of model-specific LR statistics, yields type I error which is actually *too low*, i.e. significantly lower than the nominal 0.05 rate. This demonstrates that the true, unknown theoretical distribution of the un-weighted mean LRT statistic does not conform with the chi-square distribution with one degree-of-freedom, indicating the need for further analytical work on such distributions. This is confirmed by the distributions of all $1,000$ p-values generated by weighted and un-weighted model averaging approaches, shown in Figure 4.1*c* and 4.1*d*, respectively.

4.1.3 Comparison of the Statistical Power of Different Testing Procedures

Now, we proceed to assume that there has been an association between the covariate and hazard. These associations can be linear or nonlinear. In simulations we now compare statistical power yielded by different testing procedures. The power is determined by computing the proportion of samples in which the test statistic

exceeded the critical value for the χ^2 test at $\alpha = 0.05$. In the case of linear association (Equation 3.9 Section 3.2.2), results are shown in Table 4.3 in the same order/pattern as in Table 4.1. First two columns show the empirical power calculated based on 1000 simulations for varying sample sizes ($n = 100, 300$) along with the confidence intervals. Results in subsequent columns indicate the empirical power of the AIC based model selection. It is clear that empirical power is considerably higher in the samples where a given model was selected as the best AIC model, and sharply increases as the sample size increases. The last two rows of Table 4.3 indicate that the overall power of the tests based on the minimum AIC models is significantly higher than the power of the test based on unselected models, as the confidence intervals for the corresponding proportions of rejection of H_0 do not overlap. It should be also noted that the empirical power of the tests based on particular non-linear models is only marginally lower than that of the "correct" linear model.

In the case of non linear associations described in Section 3.2.2, the general pattern of results was similar to that presented in Table 4.3. Table 4.4 summarizes the results of all simulations in which the covariate did affect the hazard and compares the overall empirical power of the four testing procedures. In the case of unselected testing without AIC selection, the reported proportions were calculated based on the results of $7,000$ tests (7 models $\times 1,000$ samples). Each of the three other testing procedures yields a single test statistic for each sample. Accordingly, their results are based on $1,000$ tests. Table 4.4 shows that, as expected, testing

Table 4.3: Comparison of the Empirical Power of the Two Testing Procedures in the Case of a Linear Association Between Covariate and log Hazard

Model	Empirical Power[a]			
	Unselected Model		Best AIC Model	
	n = 100	n = 300	n = 100	n = 300
Z	0.44	0.90	0.66	0.92
Z^2	0.40	0.87	0.49	0.91
Z^3	0.35	0.81	0.45	0.88
expZ	0.43	0.89	0.70	0.91
$\log Z$	0.43	0.85	0.66	0.97
$\frac{1}{Z}$	0.31	0.42	0.55	0.88
$\sqrt{Z}$	0.44	0.90	0.58	0.95
Overall	0.402	0.806	0.571	0.929
95% C.I.[b]	0.370-0.430	0.786-0.834	0.540-0.602	0.913-0.945

[a] Proportion of Samples in Which H_0 of No Association was Rejected Based on LR test with $\alpha = 0.05$

[b] Confidence Interval

Table 4.4: Comparison of the Empirical Power of the four Testing Procedures For Different Sample Sizes and Forms of Dose-Response Curve ($\alpha = 0.05$)

"True" Dose-Response Curve	n	Empirical Power[a]			
				Model Averaging	
		Unselected Model	Best AIC Model	Weighted	Un-weighted
Linear: $(Z - 0.5)$	100	0.520 (0.489-0.551)	0.672 (0.643-0.701)	0.600 (0.570-0.630)	0.535 (0.504-0.566)
	300	0.870 (0.849-0.891)	0.982 (0.974-0.990)	0.980 (0.970-0.999)	0.963 (0.951-0.975)
Quadratic: $2(Z - 0.5)^2$	100	0.070 (0.054-0.086)	0.215 (0.190-0.240)	0.130 (0.110-0.150)	0.037 (0.025-0.049)
	300	0.098 (0.081-0.119)	0.323 (0.294-0.352)	0.190 (0.160-0.210)	0.045 (0.032-0.058)
Quadratic: $3(Z - 0.5)^2$	100	0.090 (0.072-0.108)	0.284 (0.256-0.312)	0.160 (0.140-0.180)	0.045 (0.030-0.056)
	300	0.150 (0.128-0.172)	0.531 (0.500-0.562)	0.330 (0.300-0.360)	0.077 (0.060-0.094)
Quadratic: $4(Z - 0.5)^2$	100	0.110 (0.091-0.129)	0.392 (0.362-0.422)	0.240 (0.210-0.270)	0.045 (0.032-0.088)
	300	0.220 (0.194-0.246)	0.733 (0.706-0.760)	0.540 (0.510-0.570)	0.123 (0.103-0.143)
Threshold: Linear Spline	100	0.870 (0.849-0.891)	1.000 (1.000-1.000)	1.000 (1.000-1.000)	0.990 (0.984-0.996)
	300	0.910 (0.892-0.928)	1.000 (1.000-1.000)	0.963 1.000-1.000	1.000 (1.000-1.000)

[a] Point Estimate of the proportion of samples when H_0 is rejected at the nominal significance level $\alpha = 0.05$, with the 95% confidence interval in brackets.

the hypothesis of no association based on the minimum AIC model (2nd column) systematically increases statistical power compared to the test based on unselected Model (1st column). However, this comparison is not valid because, as demonstrated in Table 4.1, the size of the test based on best AIC model is substantially inflated. In this context, it is interesting to assess the empirical power of the modified testing procedures proposed in Section 3.2.4, that reduce the impact of *a posteriori* model selection on type I error, due to model averaging. The second last column of Table 4.4 shows that the weighted version of model averaging yields typically power that is substantially better than that of the un-selected test and only slightly lower than that of the best AIC-based procedure.

In summary, the results of our simulations confirm that the inflation of type I error due to *a posteriori* model selection creates a serious problem in the context of selecting the appropriate functional form of the dose-response relationship. In fact, the actual type I error rates may be 3 times higher than the nominal significance level even if the "best fitting" model is chosen, based on criteria such as AIC, from a set of less than ten candidate models (Table 4.1). Our results also show that a simple adjustment based on model averaging may substantially reduce the magnitude of type I error inflation (Table 4.2), although the exact distributions of the resulting test statistics remain to be identified through further numerical and/or analytical work. In view of our results, the model averaging approach which assigns Akaike weights to competing models, based on their fit to data, seems to offer an interesting

trade-off between the type I error rate, that is only slightly inflated, and empirical power that is substantially higher than in unselected testing.

4.2 Real Life Illustration

In epidemiology, prognostic models that identify risk factors for various diseases are one of the main investigative tools. Coronary heart disease (CHD) is a major cause of mortality and morbidity in Western societies. To control the incidence of CHD, several interventions and clinical guidelines have been developed by changing the levels of modifiable risk factors. In general, risks are calculated corresponding to different levels of risk factor, to predict the effect of an intervention. The validity of such predictions depends on the accuracy of the estimation of the dose-response functions describing how the risks change depending on the level of different continuous risk factors (Abrahamowicz *et al.*, 1997). This provides a practically important setting to illustrate problems investigated in our simulations.

4.2.1 Data Description

In this section, we present a secondary analysis of the public use data provided by Lipid Research Clinics (LRC) Program Prevalence and Follow-up Studies (1972-1976). The analysis is restricted to men who did not take lipid-lowering medications and the resulting data set includes $2,512$ individuals (aged between 29 to 89). The median follow-up time was reported as 12.6 years (with interquartile range, 1.2 years). The outcome of interest is the CHD death. During the follow-up 94 CHD

deaths occurred in this data set. Our main analysis focus on the effects of coronary heart disease mortality of male participants. Nine standard prognostic factors for CHD were investigated: age, systolic blood pressure (SBP), body mass index (BMI), total serum cholesterol level (TC), high density lepoprotein (HDL), smoking status (SMK), glucose intolerance (GLU), history of CHD (DEFCHD) and treatment of blood pressure with medication (BPMED). The last four factors are binary variable (presence/absence). Details of the data set is found elsewhere (LRC study, 1974).

The adjusted and unadjusted effects of various forms of continuous risk factors like TC, BMI, HDL, ratio of TC to HDL (TC/HDL) were estimated using the parametric multivariable Cox proportional hazards regression model.

For ease in calculations, all the continuous independent variables were transformed to the interval $[1, 2]$, except BMI which was transformed to have the mean at zero. However, 1 was added to transformed BMI when necessary to avoid zero values which would prevent the use of logarithms and negative power transformations.

The purpose of this study is to empirically investigate the problems related to *a posteriori* model selection and to assess the performance of the model averaging procedures.

4.2.2 Parametric Modelling of Individual Risk Factors

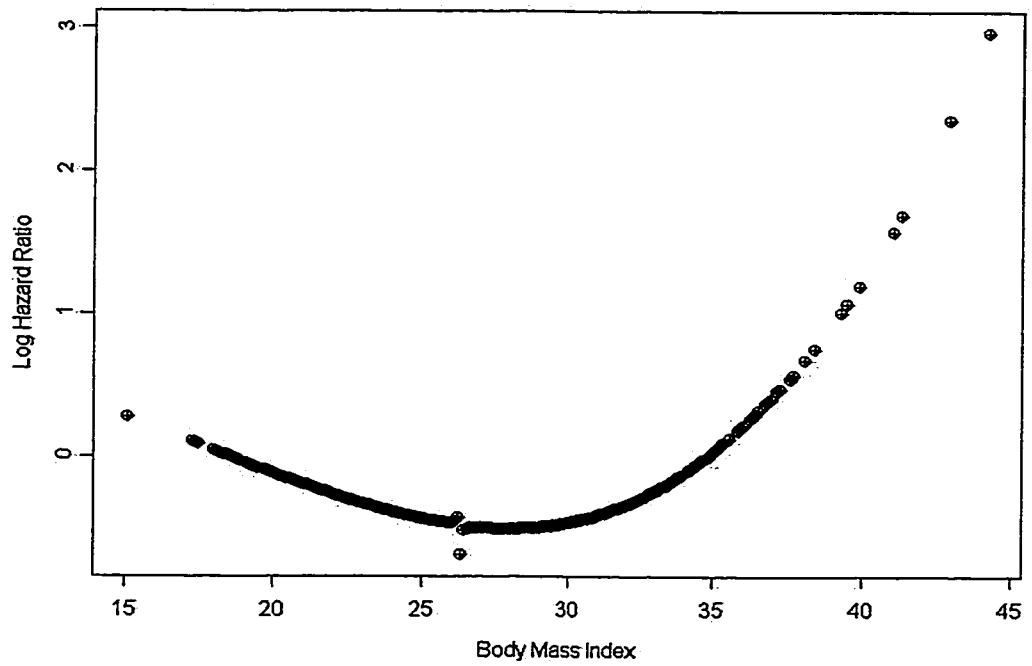
The results of separate Cox's models for the adjusted effects of TC, SBP, BMI, TC/HDL, and Age are shown in Tables 4.5 – 4.9.

For each Table, the first column gives the AIC values corresponding to different functional forms of covariate, while the second and third column give the LR tests and the corresponding p-values. The three rows at the bottom of each Table show the results of testing based on the best AIC, the weighted and unweighted testing procedures.

For most of the risk factors presented in Tables 4.5 – 4.9, all the tests are consistent in yielding the evidence of a highly significant effect on CHD mortality, regardless of the model. Thus, in those cases it won't be useful to consider other models than the linear conventional model (1st row), as long as the interest is only in establishing the statistical significance of the association. This is understandable given that TC, SBP, age or TC/HDL ratio are all very well established and very potential predictors of CHD morbidity and mortality.

By contrast, the evidence of the association between BMI and CHD mortality is less strong and in that case it may be more important to enhance the efficiency of testing the dose-response relationship. Whereas the conventional linear model (1st row) and most other models show definitely non-significant results (all p-values > 0.35), the quadratic model yields a marginally significant effect of BMI ($p < 0.10$). Figure 4.1 shows the estimated quadratic relationship between BMI and log hazard of CHD death. The shape of the curve is similar to that obtained by Abrahamowicz *et al.* (1997) using smoothing splines. Subjects in the middle range of BMI value have the lowest risk while the risks increase in both tails of the distribution. Al-

Figure 4.2: Plot of Body Mass Index (BMI) Against Log of Hazard Ratio



though the p-value is somewhat higher than the conventional cut-off of 0.05, this may be explained by a relatively low statistical power of our analyses, based on only 94 CHD deaths.

The pattern of results showed in Table 4.7 indicates that 6 out of 7 conventional parametric models considered in our study would "miss" the potentially interesting finding that there *may* be a systematic association between BMI and the risk of CHD. In other words, in order to "detect" this association, it was essential to select the quadratic transformation of BMI. The fact that the quadratic model was identified as the "best-fitting" model by the *AIC* criterion shows the potential usefulness

of this criterion in situations where the true functional form of the dose-response curve is unknown and the effect of interest is rather weak. However, as demonstrated by simulation in Section 4.1, the p-value yielded by the test based on the AIC-selected model is not valid and may considerably lower than the actual significance level. In this situation the performance of the weighted version of model averaging approach to testing is of interest as it has been show in our simulations to reduce the type I error rate inflation (Table 4.2). Table 4.7 shows that the weighted model averaging testing procedure gives a rather ambiguous result for BMI with p-value of 0.27 that is substantially lower than p-values for most parametric models but higher than p-value for the best-fitting quadratic model. It will be important to investigate the relationships between p-values yielded by different testing procedures in a broader range of situations, to see if, in selected situations, model averaging may lead to a practically meaningful change of conclusions. We believe that our results, presented in this section, provide sufficient motivation for such further endeavours.

Table 4.5: Results For the Cox PH Model With Different Functional Forms of TC

Model	AIC[a]	LR[b]	p − value
Z	1248.958	14.5113	0.0001
Z^2	1250.538	12.9318	0.0003
Z^3	1252.208	11.2616	0.0008
expZ	1251.275	12.1948	0.0005
$\log Z$	1247.492*	15.9779	0.0001
$\frac{1}{Z}$	1255.008	8.4617	0.0036
$\sqrt{Z}$	1248.210	15.2599	0.0001
Best AIC*	1247.492	15.9779	0.0001
Weighted		14.8291	0.0001
Un-Weighted		12.9427	0.0003

[a] AIC Value
[b] Likelihood Ratio for Testing the Hypothesis of No Association (LR $\sim \chi_1{}^2$ d.f.)

Table 4.6: Results For the Cox Model With Different Functional Forms of SBP

Model	AIC[a]	LR[b]	p − value
Z	1248.959	9.5145	0.0020
Z^2	1249.371	9.1016	0.0026
Z^3	1249.920	8.5534	0.0034
expZ	1249.572	9.7995	0.0028
$\log Z$	1248.674	9.9859	0.0017
$\frac{1}{Z}$	1248.487*	5.3153	0.0235
$\sqrt{Z}$	1248.802	9.6715	0.0031
Best AIC*	1248.487	5.3153	0.0235
Weighted		9.4716	0.0021
Un-Weighted		9.3611	0.0022

[a] AIC Value
[b] Likelihood Ratio for Testing the Hypothesis of No Association (LR $\sim \chi_1{}^2$ d.f.)

Table 4.7: Results of the Cox Model With Different Functional Forms of BMI

Model	AIC[a]	LR[b]	p − value
Z	1248.958	0.0301	0.8622
Z^2	1246.198*	2.7907	0.0948
Z^3	1248.140	0.8489	0.3569
ExpZ	1248.986	0.0026	0.9594
$\log Z$	1248.844	0.1442	0.7041
$\frac{1}{Z}$	1248.956	0.0322	0.8576
$\sqrt{Z}$	1248.912	0.0763	0.7823
Best AIC*	1246.198	2.7907	0.0948
Weighted		1.2008	0.2732
Un-Weighted		0.5607	0.4540

[a] AIC Value

[b] Likelihood Ratio for Testing the Hypothesis of No Association (LR $\sim \chi_1{}^2$ d.f.)

Table 4.8: Results of the Cox Model With Different Functional Forms of TC/HDL

Model	AIC[a]	LR[b]	p − value
Z	1250.357	24.2154	0.0000
Z^2	1251.902	22.6701	0.0000
Z^3	1253.768	20.8046	0.0000
expZ	1249.138	25.4345	0.0000
$\log Z$	1249.498	26.3534	0.0000
$\frac{1}{Z}$	1248.219*	24.8646	0.0000
$\sqrt{Z}$	1248.219	24.8646	0.0000
Best AIC*	1247.884	55.8964	0.0000
Weighted		25.2525	0.0000
Un-Weighted		24.2539	0.0000

[a] AIC Value

[b] Likelihood Ratio for Testing the Hypothesis of No Association (LR $\sim \chi_1{}^2$ d.f.)

Table 4.9: Results of the Cox Model With Different Functional Form of Age

Model	AIC[a]	LR[b]	p − value
Z	1248.958	54.8223	0.0000
Z^2	1251.595	52.1854	0.0000
Z^3	1254.778	49.0026	0.0000
ExpZ	1252.907	50.8737	0.0000
$\log Z$	1246.987	56.7939	0.0000
$\frac{1}{Z}$	1245.738*	58.0429	0.0000
$\sqrt{Z}$	1247.884	55.8964	0.0000
Best AIC*	1245.884	58.0429	0.0000
Weighted		55.8964	0.0000
Un-Weighted		7.9852	0.0047

[a] AIC Value

[b] Likelihood Ratio for Testing the Hypothesis of No Association (LR $\sim$ ${\chi_1}^2$ d.f.)

Chapter 5

Conclusions

In this thesis, I considered the problem of accounting for model uncertainty in the context of inference about the parameters of Cox's model, specifically addressing the issue of uncertainty involved in selecting the *optimal* transformation of a continuous covariate. I focused on the minimum AIC approach to select *a posteriori* the *optimal* transformation of a continuous predictor. In a simulation study, I investigated the *type I error* rate and the *statistical power* of likelihood ratio (LR) tests corresponding to different approaches including the minimum AIC and a new simple procedure. An empirical example was also discussed to illustrate the methodological problem considered.

The results showed that *a posteriori* model selection based on AIC leads to inflation of *type I error* rate, indicating the presence of model selection uncertainty. Therefore, a simple approach was proposed that consisted in averaging the LR statistics of all the candidate models. Two versions of the resulting statistic were considered: the unweighted and the weighted, where the weights are the Akaike weights

assigned to the different competing models. The theoretical distributions of these statistics under the null hypothesis are unknown but, as a preliminary investigation, the empirical distributions derived from the simulations were compared to a chi-square distribution with one degree of freedom. It was found that the weighted approach is able to reduce the inflation of *type I error* halfway down compared to the AIC based model selection approach. The un-weighted test also yielded *type I error* rate lower than the test based on the minimum AIC model, but still too high, which does not support the chi-square distribution with one degree of freedom of the test statistic considered. This result indicates the importance of further analytical and numerical investigations in this regard. The results also showed that the proposed weighted version of model averaging has an empirical power slightly lower than that of the best AIC-based procedure, but better than that of the unselected test procedure. In view of these results, the proposed testing procedure seems to offer an interesting trade-off between the *type I error* rate and *empirical power*

The analysis of real data on coronary heart disease illustrated the usefulness of the Akaike based approach in the situations where the true functional form of the dose-response curve is rather unknown and the association of interest is weak. This analysis of real data confirmed also the need of further investigation on the theoretical distribution of the test statistic yielded by the proposed weighted averaging procedure.

Bibliography

[1] Abrahamowicz, M., Berger, D. R., and Grover, S.A.(1997) Flexible Modelling of the Effects of Serum Cholesterol on Coronary Heart Disease Mortality. *American Journal of Epidemiology*, **145**, 8, 714-729.

[2] Abrahamowicz, M., Ciampi, A. (1991) Information Theoretic Criterion in Non-parametric Density Estimation: Bias and Variance in the Infinite Dimensional Case. *Computational Statistics and Data Analysis*, **21**, 239-247.

[3] Abrahamowicz, M., Ciampi, A., and Ramsay, J.O.(1992) Non-parametric Density Estimation for Censored Survival Data: Regression-Spline Approach. *Canadian Journal of Statistics*, **20**, 171-185.

[4] Abrahamowicz, M., Mackenzie, T., and Esdaile, J.K.(1996) Time-Dependent Hazard Ratio:Modelling and Hypothesis Testing with Application in Lupus Nephritis. *Journal of American Statistical Association*, **91**, 1432-9.

[5] Akaike, H. (1973) Information Theory as an Extension of The Maximum Likelihood Principle. *Second International Symposium on Information Theory* (Eds: B.N. Petrov, and F. Csaki). Akademiai Kiado, Budapest, 267-281.

[6] Akaike, H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control AC*, **19**, 716-723.

[7] Akaike, H. (1977) On Entropy Maximization Principle. P.R. Krishaiah (ed.), Applications of Statistics. *Biometrica*, 27-41.

[8] Akaike, H. (1978a) A new look the Bayes Procedure. *Biometrica*, **65**, 53-59.

[9] Akaike, H. (1978b) A Bayesian Analysis of the Minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, **30**, 9-14.

[10] Akaike, H. (1979) A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting. *Biometrica*, **66**, 237-242.

[11] Akaike, H. (1980) Likelihood and the Bayes Procedure (with discussion. In J. M. Bernardo, M.H. De Groot, D.V. Lindley, and A.F.M. Smith (eds.) *Bayesian Statistics*, University Press, Valencia, Spain, 143-203.

[12] Akaike, H. (1981a) Likelihood of a Model and Information Criteria. *Journal of Econometrics*, **16**, 3-14.

[13] Akaike, H. (1981b) Modern Development of Statistical Methods. In P. Eykhoff (ed.). *Trends and Progress in System Identification*. Pergamon Press, Paris, 169-184.

[14] Akaike, H. (1983b) Information Measures and Model Selection. *International Statistical Institute*, **44**, 277-291.

[15] Akaike, H. (1987) Factor Analysis and AIC. *Psychometricka*, **52**, 317-332.

[16] Amari, S. (1993) Mathematical Methods of Neurocomputing (Eds: O.E. Bandorff-Nielson, J.L. Jensen, and W.S. Kendall). *Networks and Chaos-Statistical and Probabilistic Aspects*. Chapman and Hall, New York.

[17] Altman, D. G. (1993) Categorizing Continuous Variable. *British Journal of Cancer*, **64**, 975.

[18] Altman, D. G., Anderson, P. K. (1989) Bootstrap Investigation of the Stability of a Cox's Regression Model. *Statistics in Medicine*, **8**, 771-783.

[19] Atkinson, R. (1985) Plots, Transformation and Regression, *Oxford: Oxfrod Scientific.*

[20] Atilgan, T. (1996) Selection of Dimension and Basis for Density Estimation and Selection of Dimension, Basis and Error Distribution of Wildlife. *Journal of Applied Ecology*, **33**, 339-347.

[21] Azzalini, A. (1996) *Statistical Inference-Based on the Likelihood.*Chapman and Hall, London.

[22] Bancroft, T.A., and Han, C.P. (1977) Inference Based on Conditional Specification:A note and a Bibliography. *International Statistical Review*, **45**, 117-281.

[23] Berger, J. O., and Wolpert, R. L. (1984) The Likelihood Principle. *Institute of Mathematical Statistics Monograph*, **6**.

[24] Boucher, K.M., Slattery, M.L., Berry, T.D., Quesenberry, C., and Anderson, K. (E1998) Statistical Methods in Epidemiology: A Comparison of Statistical Methods to Analyze Dose-Response and Trend Analysis in Epidemiologic Studies.*Journal of Clinical Epidemiology*, **51**, 12, 1223-1233.

[25] Bozdogan, H. (1987) Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. *Psychometrika*, **52** 345-370.

[26] Bozdogan, H. (1988) A New Model-Selection Criterion (Eds: H.H. Bock), *Classification and Related Methods of Data Analysis.* North-Holland Publishing Company, Amsterdam, 599-608.

[27] Box, G.E.P., and Jenkins, G.M. (1970) *Time Series Analysis: Forecasting and Control.*Holden-Day, London, **17**.

[28] Breiman, L. (1988) Sub model Selection and Evaluation in Regression. *The Conditional Case and Little Bootstrap, Technical Report* **169**, University of California, Berkeley, Department of Statistics.

[29] Breiman, L. (1992) The Little bootstrap and other Methods for dimensionality in regression: X-fixed Prediction Error.*Journal of The American Statistical Association*, **24**.

[30] Breiman, L. (1996) Heuristics of Instability and Stabilization in Model Selection.*The Annals of Statistics*, **24**, 2350-2383.

[31] Brenner, H., and Blettner, M. (1997) Controlling for Continuous Confounders in Epidemiologic Research.*Epidemiology*, **8**, 429-434.

[32] Breslow, N.E. and Day, N.E. (1980) Statistical Methods in Cancer Research. The Analysis of Case-Control Studies. Lyon, France: *International Agency for Research on Cancer*, (IARC Scientific Publication No. 32), **1**.

[33] Breslow N.E., and Day, N.E. (1987) Statistical Methods in Cancer Research. The Design and Analysis of cohort studies. Lyon, France: *International Agency for Research on Cancer*, (IARC Scientific Publication no. 82), **2**.

[34] Brown, C.C., Kipnis, V., Freedman, L.S., Hartman, A.M., Schatzkin, A., and Watcholder, S. (1994) Energy Adjustment Methods for Nutritional Epidemiology: The Effect of Categorization.*American Journal of Epidemiology*, **139**, 323-338.

[35] Buckland, S.T., Burnham, K.P., and Augustin, N.H. (1997) Model Selection: An Integral Part of Inference. *Biometrics*, **53**, 603-618.

Wildlife. *Journal of Applied Ecology*, **30**, 478-495.

[36] Burnham, K.P., and Anderson, D.R. (1998) *Model Selection And Inference: A Practical Information Theoretic Approach.* Springer-Verlag.

[37] Burnham, K.P., Anderson, D.R., and White, G.C. (1995) Selection Among Open Population Capture-Recapture Data. *Biometrics*, **51**, 888-898.

[38] Burman, P. (1989) A Comparative Study of Ordinary Cross-Validation, V-hold Cross-Validation and Repeated Learning-Testing Methods. *Biometrica*, **76**, 503-514.

[39] Burnham, K.P., Anderson, D.R., and Whilte, G.C. (1994) Evaluation of the Kullback-Liebler Discrepancy for Model Selection in One Population Capture-Recapture models. *Biometrical Journal*, **51**, 888-898.

[40] Carlin, B.P., and Chib, S. (1995) Bayesian Model Choice Via Markov Chain Monte Carlo Methods.*Journal of The Royal Statistical Society*, Series B **57**, 473-484.

[41] Chatfield, C. (1995) Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society*, Series A **158**, 419-466.

[42] Cox, D.R. (1958) The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society*, Series B, **20**, 215-242.

[43] Cox, D.R. (1972) Regression Models and Life Tables. *Journal of the Royal Statistical Society*, Series B, **34**, 187-220.

[44] Cox, D.R., and Oakes, D. (1984) *Analysis of Survival Data.* Chapman and Hall, London.

[45] Craven, P., and Wahba, G. (1979) Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross Validation. *Numerical Mathematics*, **31**, 377-403.

[46] Draper, N.R., Smith, H. (1981) *Applied Regression Analysis (2nd ed.)*, New York, John Wiley.

[47] Draper, D. (1995) Assessment and Propagation of Model Uncertainty (with discussion).*Journal of Royal Statistical Society*, Series B; **57**, 45-97.

[48] Edwards, A.W.F. (1992) *An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference.* Cambridge University Press, London.

[49] Efron, B., and Tibshirani R.J. (1993) *An Introduction to Bootstrap.* New York, Chapman and Hall.

[50] Elston, R.C. and Johnson, W.D. (1994) *Essentials of Biostatistics*, F.A. Davis (ed.).

[51] Forthofer, R.N. (1995) *Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery.* Academic Press.

[52] Flemming, T.R., and Harrington, D.P. (1991) *Counting Process and Survival Analysis*, John Wiley and Sons, New York.

[53] Geisser, S. (1975) The Predictive Sample Reuse Method with Approaches to Calculating Marginal Densities.*Journal of The American Statistical Association*, **85**.

[54] Gilchrist, W. (1984) *Statistical Modelling*, John Wiley and Sons, New York.

[55] Guo, S., Roche, A.F., and Moore, W.M. (1988) Reference Data For Head Circumference and 1-month increments, From 1 to 14 month of Age. *Journal of Pediatrics*, **113**, 490-494.

[56] Goutis, C.W.J., and Casella, G. (1995) Frequentist Post-Data Inference. International Statistical Theories and the Use of Model Selection Criteria. *Journal of Econometrics*, **67**, 173-187.

[57] Gray, R.J. (1992) Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. *Journal of the American Statistical Association*, **87**, 942-951.

[58] Greenland, S. (1995) Dose-Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis. *Epidemiology*, **6**, 356-365.

[59] Halpern, J. (1982) Maximally Selected Chi-Square Statistics For Small Samples. *Biometrics*, **38**, 1017.

[60] Hannan, E.J., and Quinn, B.G. (1979) The Determination of the Order of an Auto-regression. *Journal of The Royal Statistical Society*, Series B, **41**, 190-195.

[61] Harrell, F.E., Jr., Lee, K.L., and Pollock, B.G. (1988) Regression Model in Clinical Studies: Determining Relationships Between Predictors and Response. *Journal of National Cancer Institute*, **80**, 1198-1202.

[62] Harvey, A.C.(1981) *Time Series Models*, Phillip Allen, Oxford.

[63] Hastie, T., and Tibshirani, R. (1986) Generalized Additive Models (With Discussion). *Statistical Science*, **1**,297-318.

[64] Hastie, T., and Tibshirani, R. (1987) Generalized Additive Models: Some Applications.*Journal of the American Statistical Association*, **82**,371-386.

[65] Hastie, T., Tibshirani, R. (1990) *Generalized Additive Models.*Chapman and Hall, New York.

[66] Hill, A. B. (1965) The environment and disease: Association or causation? *Proc R Soc Med*, **58**, 295-300.

[67] Hjorth, J.S.U. (1994) *Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap.* Chapman and Hall, London.

[68] Hosmer, D.W., and Lemeshow, S. (1989) *Applied Logistic Regression.* John Wiley and Sons, New York.

[69] Hurvich, C.M., and Tsai, C.L. (1989) Regression and Time Series Model Selection in Small Samples.*Biometrika*, **78**, 499-509.

[70] Hurvich, C.M., and Tsai, C.L. (1990) The Impact of Model Selection on Inference in Linear Regression. *The American Statistician*, **44**, 214-217.

[71] Isaacs, D., Altman, D.G., Tidemarsh, C.E., Valmer, H.B., and Webster, A.D.B. (1983) Serum Immunoglobin Concentration in Preschool Children Measured by Laser Nephelometry: Reference Ranges For IgG, IgA, IgM. *Journal of Clinical Pathology*, **36**, 1193-1196.

[72] Kabaila, P. (1995) The Effect of Model Selection on Confidence Regions and Prediction Regions. *Econometric Theory*, **11**, 537-549.

[73] Kalbfleisch, J.D., and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data*.John Wiley and Sons, New York.

[74] Kapur, J.N. and Kesavan, H.K. (1992) *Entropy Optimization Principles with Applications*. Academic Press, London.

[75] Kishino, H.H., Kato, H., Kasamatsu, F., and Fujise, Y. (1991) Detection of Heterogeneity and Estimation of Population Characteristics from Field Survey Data: 1987/88 Japanese Feasibility Study of The Southern Hemisphere Minke Whales.*Annals of the Institute of Statistical Mathemtaics*, **43**, 435-453.

[76] Kleinbaum, D.G., Kupper, L., and Muller, K.E. (1988) Applied Regression Analysis and Other Multivariable Methods, 2nd edition, Boston: PWC-Kent Publishing Company.

[77] Kooperberg, C., Stone, C.J., and Truong, Y.K. (1995) Hazard Regression. *Journal of the American Statistical Association*, **90**, 78-94.

[78] Kuk, A.Y.C. (1984) All Subsets Regression in a Proportional Hazards Model. *Biometrika*, **71**, 587-592.

[79] Kullback, S., and Liebler, R. A. (1951) On Information and Sufficiency. *Annals of Mathematical Statistics*, **22**, 79-86.

[80] Kullback, S. (1959) *Information Theory and Statistics.* John Wiley and Sons, New York.

[81] Laud, P.W., and Ibrahim, J.G. (1995) Predictive Model Selection. *Journal of The Royal Statistical Society*, Series B, **57**.

[82] Lebreton, J.D., Burnham, K.P., Clobert, J., and Anderson, D. R. (1992) Modelling Survival and Testing Biological Hypothesis Using Marked Animals: A Unified Approach with Case Studies.*Ecological Monograph*, **62**, 67-118.

[83] Lehmann, E.L. (1983) *Theory of Point Estimation.* John Wiley and Sons, New York.

[84] Linhart, H., and Zucchini, W. (1986) *Model Selection.* John Wiley and Sons, New York.

[85] (1974) Lipid Research Clinics Program. *Manual of Laboratory Operations, National Institutes of Health.*NIH Publications No. 75-628, **1**.

[86] Maclure, M., and Greenland, S. (1992) Tests for Trend and Dose-Response: Misinterpretations and Alternatives.*American Journal of Epidemiology*, **135**, 96-104.

[87] Madigan, D., and Raftery, A.E. (1994) Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window.*Journal of The American Statistical Association*, **89**, 1535-1546.

[88] Mallows, C. L. (1973) Some Comments on C_p. *Technometrics*, **12**, 591-612.

[89] Mallows, C. L. (1995) Some Comments on C_p. *Technometrics*, **37**, 362-372.

[90] McCullagh, P., and Pregibone, D. (1985) Discussion Comments On the Paper by Diaconis and Efron. *Annals of Statistics*, **13**, 898-900.

[91] McCullagh, P., and Nelder, J.A. (1989) *Generalized Linear Models*. 2nd Ed., Chapman and Hall, New York.

[92] Miller, A. J. (1984) *Subset Selection in Regression*. Chapman and Hall, New York.

[93] Mooney, C. Z., and Duval, R.D. (1993) *Bootstrapping: A Non-parametric Approach to Statistical Inference*, Sage Publications, London.

[94] Moore, D.F. (1987) Modelling the Extraneous Variance in The Presence of Extra-Binomial Variation. *Journal of The Royal Statistical Society*, **36**, 8-14.

[95] Mosteller, F., and Tukey, J.W. (1968) Data Analysis, Including Statistics. (Eds: G. Lindzey, and E. Aronson). *Handbook of Social Psychology*, Vol. 2. Addison-Wesley, Reading, MA.

[96] Neter, J., Wasserman, W., and Kutner, M.H., (1990) Applied Linear Statistical Models (3rd ed.). Irwin.

[97] Newman, K. (1997) Bayesian Averaging of generalized linear models for passive integrated trans ponder tag recoveries from salmonids in the Snake River.*North American Journal of Fisheries Management*, **17**, 362-377.

[98] O'Sullivan, F. (1988) Nonparamteric Estimation of Relative Risk Using Spline and Cross Validation.*SIAM Journal on Scientific and Statistical Computing*, **9**, 531-542.

[99] Potscher, B.M. (1991) Effects of Model Selection on Inference. *Econometric Theory*, **7**, 163-185.

[100] Quantin, C., Abrahamowicz, M., Moreau, T., Bartlett, G., Mackenzie, T., Tazi, M, A., Lalonde, L., and Faivre, J. (1999) Variation Over Time of the Effects of Prognostic Factors in a Population-based Study of Colon Cancer: Comparison of Statistical Models. *American Journal of Epidemiology*, **150**, 11, 1188-1200.

[101] Raftery, A.E., Madigan, D., and Hoeting, J. (1993) Model Selection and Accounting for Model Uncertainty in Linear Regression Models. Technical Report No. **262**, Department of Statistics, University of Washington, Seattle.

[102] Raftery, A.E., and Madigan, D. (1994) Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window.*Journal of The American Statistical Association*, **89**, 1535-1546.

[103] Raftery, A.E., Madigan, D., and Volinsky C.T. (1995) Accounting For Model Uncertainty in Survival Analysis Improves Predictive Performance (with Discussion).In J. Bernardo, J. Berger, J., A. David, and A. Smith (Eds.). *Bayesian Statistics*, **5**, pp 323-349, Oxford University Press.

[104] Ramsay, J.O. (1988) Monotone Regression Splines in Action: with Discussion. *Statistical Science*, **3**, 425-461.

[105] Ramsay, J.O., Abrahamowicz, M. (1989) Binomial Regression With Monotone Splines: A Psychometric Application. *Journal of the American Statistical Association*, **84**, 916-25.

[106] Rawlings, J. O. (1988) *Applied Regression Analysis: A Research Tool.*Wadsworth, Inc., Belmont, CA.

[107] Ross, S, M. (1997) *Simulation.* 2nd edition, Academic Press, New York.

[108] Royston, P., and Altman, D.G. (1994) Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling.*Applied Statistics*, **43**, 429-467.

[109] Royall, R. M. (1997) *Statistical Evidence: A Likelihood Paradigm.* Chapman and Hall, London.

[110] Rissanen, J. (1989) Stochastic Complexity in Statistical Inquiry. *World Scientific*, Series in Computer Science, **15**.

[111] Rencher, A.C., and Pun, F.C. (1980) Inflation of R^2 in Best subset Regression. *Technometrics*, **22**, 49-53.

[112] Rothman, K.J. (1986) *Modern Epidemiology.* Boston: Little, Brown.

[113] Saber, G.A.F. (1977) *Linear Regression Analysis.* John Wiley and Sons, New York.

[114] Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986) Akaike Information Criterion Statistics. *KTK Scientific Publishers*, Tokyo.

[115] Sauerbrei, W. and Schumacher, M. (1992) A Bootstrap Re-Sampling Procedure For Model Building: Application To The Cox's Regression Model. *Statistics in Medicine*, **11**, 2093-2109.

[116] Sauerbrei, W. and Royston, P. (1999) Building Multivariable Prognostic and Diagnostic Models: Transformation of The Predictors Using Fractional Polynomials. *Journal of The Royal Statistical Society*, Series A, **162**, 71-94.

[117] Schulgen, G., Lausen B, Olsen, J.H., and Schumacher, M. (1994) Outcome-Oriented Cut-Points in Analysis of Quantitative Exposure. *American Journal of Epidemiology*, **140**,172-184.

[118] Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461-464.

[119] Sclove, S.L. (1987) Application of Some Model-Selection Criterion to Some Problems in Multivariate Analysis. *Psychometrika*, **52**, 333-343.

[120] Sclove, S.L. (1994) Small-sample and Large Sample Statistical Model Selection Criteria. In P. Cheeseman, and R. W. Oldford (eds.). *Selecting Models From Data*, (Eds: P. Cheeseman, and R. W. Oldford). Springer-Verlag, New York.

[121] Shao, J., and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, New York.

[122] Shibata, R. (1983) *A Theoretical View of The Use of AIC: From Data to Model*, (Eds: O.D. Anderson). Springer-Verlag, London.

[123] Shibata, R. (1986) *Consistency of Model Selection and Parameter Estimation: Essays in Time Series and Allied Processes*, (Eds: J. Gani, and M.B. Priestly). *Journal of Applied Probability*, Special Volume **23A**.

[124] Shibata, R. (1989) *Statistical Aspects of Model Selection: From Data To Model*, (Eds: J.C. Willems). Springer-Verlag, London.

[125] Sleeper, L.A., and Harrington, D.P. (1990) Regression Splines in the Cox Model With Application to Covariate Effects in Liver Disease. *Journal of The American Statistical Association*, **85**, 941-949.

[126] Snedecor, G.W., and Cochran, W.G. (1967) *Statistical Methods*, 6th Ed., Iowa, Iowa State University Press.

[127] Sobol, I.M. (1994) *A Primer for the Monte Carlo Method.* CRC Press.

[128] S-PLUS (1995) S-PLUS Version 4.0 Seattle, WA: MathSoft.

[129] Stone, M. (1974) Cross-validatory Choice and Assessment of Statistical Predictions (With Discussion). *Journal of The Royal Statistical Society*, Series B **39**, 111-147.

[130] Stone, M. (1977) An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion.*Journal of The Royal Statistical Society*, Series B **39**, 44-47.

[131] Stone, C.J (1985) Additive Regression and Other Nonparamteric Models.*Annals of Statistics*, **13**, 689-705.

[132] Suguira, N., (1978) Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections.*Communications in Statistics: Theory and Methods.* **A7**, 13-26.

[133] Volinsky, C.T. , (1997) *Bayesian Model Averaging for Censored Survival Models.* Unpublished Ph.D. Thesis, University of Washington.

[134] Vacek, P.A. (1997) Assessing the Effect of Intensity When Exposure Varies Over Time. *Statistics In Medicine*, **16**, 505-513.

[135] Wedderburn, R. W. M. (1974) Quasi-Likelihood Functions, Generalized Linear Models, and The Gauss-Newton Method. *Biometrika*,**61**, 439-447.

[136] Wegman, E.J., and Wright, J.W., (1983) Spline in Statistics. *Journal of The American Statistical Association*, **78**, 351-366.

[137] Weinberg, C.R. (1995) How Bad is Categorization? (Editorial), *Epidemiology* **6**, 345-347.

[138] Westfall, P.H., and Young, S.S. (1993) *Re-sampling-Based Multiple Testing: Examples and Methods For P-Value Adjustment.* John Wiley and Sons, New York.

[139] Williams, D.A. (1982) Extra-Binomial Variation in Logistic Linear Models. *Applied Statistics*, **31**, 144-148.

[140] Wold, S. (1974) Spline Functions In Data Analysis. *Technometrics*, **16**, 1-11.

[141] Zhang, P. (1993a) Model Selection Via Multi-Fold Cross-Validation. *Annals of Statistics*, **20**, 299-313.

[142] Zhang, P. (1994) *On the Choice of Penalty Term in Generalized FPE Criterion: Selecting Models From Data* (Eds: P. Cheeseman, and R.W. Oldford). Springer-Verlag, New York.

[143] Zhao L.P., and Kolonel, L.N. (1992) Efficiency Loss From Categorizing Quantitative Exposures into Qualititative Exposures in Case-Control Studies.*American Journal of Epidemiology*, **136**, 464-74.

Appendix A

S-PLUS Code For The Simulation Study

A.1 Simulation for the Unselected Models

```
# Hazard function doesn't depend on Covariate
# To generate 1000 samples with sample size n=100,300.
nsims <- 1000
n <- 100
aicmat <- matrix(data=0, nrow=nsims, ncol=7)
pvalmat <- matrix(data=0, nrow=nsims, ncol=7)
set.seed(9949149)
for (i in 1:nsims) {
res <- aic.f(n)
aicmat[i,] <- res$aic
        pvalmat[i,] <- res$pval
        }
Status <- function(z, y) {
ifelse(z > y, 0, 1)
}
Time <- function(z, y)  {
ifelse(z > y, y, z)
}
# Function to return AIC and P-VALUE
aic.f <- function(n) {
u <- runif(n)
```

```
tm <- (-0.5)*log(u)  # t = Failure Time
st <- runif(n) # st=Survivor Function
ct <- (-1/1.5)*log(s)  # ct = Censoring time generated from exp(1.5)
di <- Status(tm,ct)   # Storing the status variable
fi <- Time(tm,ct)   # Storing the observed survival time variable
# Generating 7 different functions of covariates
x <- runif(n)        # x = Covariate values
x2 <- x^2
x3 <- x^3
x4 <- exp(x)
x5 <- log(x)
x6 <- 1/x
      x7 <- sqrt(x)
# Storing the survival data
survdat <- data.frame(fi,di,x,x^2,x^3,exp(x),log(x),1/x,sqrt(x))
# Fitting the Cox PH model for the all 7 forms of covariates
      coxfit1 <- coxph(Surv(fi, di) ~ x, survdat)
coxfit2 <- coxph(Surv(fi, di) ~ x^2, survdat)
coxfit3 <- coxph(Surv(fi, di) ~ x^3, survdat)
coxfit4 <- coxph(Surv(fi, di) ~ exp(x), survdat)
coxfit5 <- coxph(Surv(fi, di) ~ log(x), survdat)
coxfit6 <-coxph(Surv(fi, di) ~ I(1/x), survdat)  # Note I(1/x)
coxfit7 <-coxph(Surv(fi, di) ~ sqrt(x), survdat)
# Calculating AIC for each of the 7 models
aic <- c(-2*coxfit1$loglik[2] + (2 * 1), -2*coxfit2$loglik[2] + (2 * 1),
         -2*coxfit3$loglik[2] + (2 * 1), -2*coxfit4$loglik[2] + (2 * 1),
         -2*coxfit5$loglik[2] + (2 * 1), -2*coxfit6$loglik[2] + (2 * 1),
         -2*coxfit7$loglik[2] + (2 * 1))
# Calculating Likelihood Ratio Statistics(LRT) for each of the 7 models
 <- 2*c((coxfit1$loglik[2]-coxfit1$loglik[1]),(coxfit2$loglik[2]
       coxfit2$loglik[1]),(coxfit3$loglik[2]-coxfit3$loglik[1]),
       (coxfit4$loglik[2]-coxfit4$loglik[1]),(coxfit5$loglik[2]-
       (coxfit7$loglik[2]-coxfit7$loglik[1]))
pval <- (1-pchisq(lrt,c(1,1,1,1,1,1,1)))
return(list("aic"=aic,"pval"=pval,"lrt"=lrt))
}
# To find the frequency distribution of the p-values for 20 intervals
```

```
p <- round(pvalmat,2)
pgroup1 <- cut(p[,1], breaks=c(0,0.05,0.10,0.15,0.20,0.25,0.30,0.35,
0.40,0.45,0.50,0.55,0.60,0.65,0.70,0.75,0.80,0.85,0.90,0.95,1.0)
,include.lowest=T)   # same for 20 intervals
# Storing the values for individual models
m1<-cbind(table(pgroup1))
m2<-cbind(table(pgroup2))
m3<-cbind(table(pgroup3))
m4<-cbind(table(pgroup4))
m5<-cbind(table(pgroup5))
m6<-cbind(table(pgroup6))
m7<-cbind(table(pgroup7))
freq.table<-cbind(m1,m2,m3,m4,m5,m6,m7)
allfreq <- cbind(freq.table,apply(freq.table,1,sum))

# To calculate the percentage distribution of p values
perct.table1<-round((cbind(freq.table[,1]))/sum(cbind(freq.table[,1],)),2)
perct.table2<-round((cbind(freq.table[,2]))/sum(cbind(freq.table[,2],)),2)
perct.table3<-round((cbind(freq.table[,3]))/sum(cbind(freq.table[,3],)),2)
perct.table4<-round((cbind(freq.table[,4]))/sum(cbind(freq.table[,4],)),2)
perct.table5<-round((cbind(freq.table[,5]))/sum(cbind(freq.table[,5],)),2)
perct.table6<-round((cbind(freq.table[,6]))/sum(cbind(freq.table[,6],)),2)
perct.table7<-round((cbind(freq.table[,7]))/sum(cbind(freq.table[,7],)),2)
perct.table <- cbind(perct.table1,perct.table2,perct.table3,perct.table4,
perct.table5,perct.table6,perct.table7)
# To calculate the percentage distribution of p values according to
the group(overall)
finperct.table <- apply(freq.table,1,sum)
finpct <- matrix(data=0,nrow=20,ncol=1)
for (i in 1:20) {
finpct[i,] <- (cbind(finperct.table)[i,])/sum(finperct.table)
                }
finpct <- round(cbind(perct.table, finpct),2)
# To find the 95% C.I. for overall proportion
CITAB <- matrix(data=0,nrow=20,ncol=2)
for (i in 1:20) {CITAB[i,] <-
(cbind ((finpct[i,8]-sqrt((1-finpct[i,8])*(finpct[i,8])/nsims)*1.96),
```

```
(finpct[i,8]+sqrt((1-finpct[i,8])*(finpct[i,8])/nsims)*1.96)))}
fintable <- cbind(finpct,round(CITAB,4))
dimnames(allfreq)<- list(NULL,c("m1","m2","m3","m4","m5","m6",
"m7","TOTAL"))
dimnames(fintable)<-list(NULL,c("m1","m2","m3","m4","m5","m6","m7",
"ALL","LB","UB"))
# Final Results
fintable<-round(fintable,3)
allfreq
fintable
#Percentage of censoring
sum(di)/100
```

A.2 Simulation for the Best AIC models

```
# Hazard function doesn't depend on Covariate.
# To generate 1000 samples with sample size n=100,300.
nsims <- 1000
n <- 100
aicmat <- matrix(data=0, nrow=nsims, ncol=7)
pvalmat <- matrix(data=0, nrow=nsims, ncol=7)
set.seed(9949149)
for (i in 1:nsims) {
res <- aic.f(n)
aicmat[i,] <- res$aic
        pvalmat[i,] <- res$pval
              }
# To create a vector of minimum AIC and corresponding p-values
is.min <- function(x) x==min(x)
mins <- function(x)
(1:length(x)) [is.min(x)]
index <- apply(aicmat,1,mins)
pvalaic <- index   # p values corresponds to min AIC
for (i in 1:nsims) pvalaic[i] <- pvalmat[i,index[i]]
Status <- function(z, y) {
ifelse(z > y, 0, 1)
}
```

```
Time <- function(z, y)  {
ifelse(z > y, y, z)
}
# Function to return AIC and P-VALUE
aic.f <- function(n) {
u <- runif(n)
tm <- (-0.5)*log(u) # t = Time to event
st <- runif(n)
ct <- (-1/1.5)*log(st)  # c = Censoring Time with exp(1.5)
di <- Status(tm,ct)   # Storing the status variable
fi <- Time(tm,ct)   # Storing the survival time variable
x <- runif(n)        # x = Covariate values
# Generating different covariates
x2 <- x^2
x3 <- x^3
x4 <- exp(x)
        x5 <- log(x)
x6 <- 1/x
        x7 <- sqrt(x)
# Storing the survival data
  survdat <- data.frame(fi,di,x,x^2,x^3,exp(x),log(x),1/x,sqrt(x))
# fitting the Cox PH model for the all 7 forms of covariates
      coxfit1 <- coxph(Surv(fi, di) ~ x, survdat)
coxfit2 <- coxph(Surv(fi, di) ~ x^2, survdat)
coxfit3 <- coxph(Surv(fi, di) ~ x^3, survdat)
coxfit4 <- coxph(Surv(fi, di) ~ exp(x), survdat)
coxfit5 <- coxph(Surv(fi, di) ~ log(x), survdat)
coxfit6 <-coxph(Surv(fi, di) ~ I(1/x), survdat)  # Note I(1/x)
coxfit7 <-coxph(Surv(fi, di) ~ sqrt(x), survdat)
# Calculating AIC for each of the 7 models
aic <- c(-2*coxfit1$loglik[2] + (2 * 1), -2*coxfit2$loglik[2] + (2 * 1),
         -2*coxfit3$loglik[2] + (2 * 1), -2*coxfit4$loglik[2] + (2 * 1),
         -2*coxfit5$loglik[2] + (2 * 1), -2*coxfit6$loglik[2] + (2 * 1),
        -2*coxfit7$loglik[2] + (2 * 1))
# Calculating Likelihood Ratio Statistics(LRT) for each of the 9 models
lrt <- 2*c((coxfit1$loglik[2]-coxfit1$loglik[1]),(coxfit2$loglik[2]
 -coxfit2$loglik[1]),(coxfit3$loglik[2]-coxfit3$loglik[1]),
```

```
        (coxfit4$loglik[2]-coxfit4$loglik[1]), (coxfit5$loglik[2]-
        (coxfit7$loglik[2]-coxfit7$loglik[1]))
pval <- (1-pchisq(lrt,c(1,1,1,1,1,1,1)))
return(list("aic"=aic,"pval"=pval))
}
# To find the distribution (%) of minimum AIC according to different models
freq.ind <- table(index)
perct <- round(freq.ind/nsims,2)
finout <- rbind(freq.ind, perct)
pvalfin <- signif(pvalaic,2)
# To find the distribution of the p-values
pgroup <- cut(pvalfin, breaks=c(0,0.05,0.10,0.15,0.20,0.25,0.30,0.35,0.40,
0.45,0.50,0.55,0.60,0.65,0.70,0.75,0.80,0.85,0.90,0.95,1.0),
include.lowest=T)   # 20 intervals
freq.table<-table(pgroup,index)
# To calculate the percentage distribution of p values
perct.table1<-round((cbind(freq.table[,1]))/sum(cbind(freq.table[,1],)),2)
perct.table2<-round((cbind(freq.table[,2]))/sum(cbind(freq.table[,2],)),2)
perct.table3<-round((cbind(freq.table[,3]))/sum(cbind(freq.table[,3],)),2)
perct.table4<-round((cbind(freq.table[,4]))/sum(cbind(freq.table[,4],)),2)
perct.table5<-round((cbind(freq.table[,5]))/sum(cbind(freq.table[,5],)),2)
perct.table6<-round((cbind(freq.table[,6]))/sum(cbind(freq.table[,6],)),2)
perct.table7<-round((cbind(freq.table[,7]))/sum(cbind(freq.table[,7],)),2)
perct.table <- cbind(perct.table1,perct.table2,perct.table3,perct.table4,
perct.table5,perct.table6,perct.table7)
# To calculate the percentage distribution of p values according to the
group (overall)
finperct.table <- apply(freq.table,1,sum)
finpct <- matrix(data=0,nrow=20,ncol=1)
for (i in 1:20) {
finpct[i,] <- (cbind(finperct.table)[i,])/nsims}}
finpct <- cbind(perct.table, finpct)
# To find the 95% C.I. for overall proportion
CITAB <- matrix(data=0,nrow=20,ncol=2)
for (i in 1:20) {CITAB[i,] <-
(cbind ((finpct[i,8]-sqrt((1-finpct[i,8])*(finpct[i,8])/nsims)*1.96),
(finpct[i,8]+sqrt((1-finpct[i,8])*(finpct[i,8])/nsims)*1.96)))}
```

```
fintable <- cbind(finpct,round(CITAB,4))
# Final Results
allfreq <- cbind(freq.table, totfreq <- apply(freq.table,1,sum))
dimnames(allfreq) <-list(NULL,c("m1","m2","m3","m4","m5","m6","m7",
"TOTAL"))
fintable <- round(fintable,3)
dimnames(fintable) <-list(NULL,c("m1","m2","m3","m4","m5","m6","m7",
"ALL","LB","UB"))
allfreq
fintable
finout
cat("Percentage of censoring:","\n")
sum(di)/100
\section{ Simulation for Weighted and Unweighted Averaging Model}
\begin{verbatim}
# Required additional codes to calculate the Type I Error and
Empirical Power for the Weighted and Unweighted Averaging Model
# To calculate the Akaike weights, weighted LR test and
corresponding p-values
minaic <-  min(aic)
aicdiff <- round((aic-minaic),4)
sumdiff <- sum(exp((-1/2)*(aicdiff)))
wi <-  exp((-1/2)*(aicdiff))/(sumdiff)
lrtw <- sum(lrt*wi)
pvalw <- (1-pchisq(lrtw,1))
# To claculate the un-weighted LR test and corresponding p-values
lrtuw <- sum((1/7)*(lrt))
pvaluw <- (1-pchisq(lrtuw,1))
```

A.3 Additional Code for the calculation of the Empirical Power

```
# Codes to calculate the Empirical Power when there is covariate effect
(linear \& non linear).
# Function to return AIC and P-VALUE
Status <-  function(z,y) {
   ifelse (z > y, 0, 1)
```

```
}
Time <-  function(z, y) {
ifelse$(z > y, y, z)
}
aic.f <- function(n) {
x <- runif(n)    # x=Covariate values
fx <- (x-0.5)    # A Linear function of covariate x
fx <- c*(x-0.5)^2$   # Quadratic function of covariate; c=2,3 or 4
fx <- ifelse (x<0.5,0,4*(x-0.5)) # A Non Linear (bi-spline) function
     of covariate
lmda <- exp(1+fx)  # lmda= Log HR which depends on the covariate
st <- runif(n)   # st=Exponential Survival
tm <- (-1/lmda)*log(st)  # tm = Failure Time
ct <- (-1/1.5)*log(st) # ct = Censoring  Time exp(1.5)
di <-  Status(tm,ct)       # Storing the status variable
fi <- Time(tm,ct)             # Storing the observed survival time variable
# Generating $7$ different function of a single covariate
x1 <-  x
x2 <-  x^2
x3 <-  x^3
x4 <-  exp(x)
      x5 <- log(x)
x6 <- 1/x
      x7 <- sqrt(x)
# Storing the survival data
  survdat <- data.frame(fi,di,x,x^2,x^3,exp(x),log(x),1/x,sqrt{x})
# Fitting the Cox PH model for the all 7 forms of the covariate
      coxfit1 <- coxph(Surv(fi, di) ~ x, survdat)
coxfit2 <- coxph(Surv(fi, di) ~ x^2, survdat)
coxfit3 <- coxph(Surv(fi, di) ~ x^3, survdat)
coxfit4 <- coxph(Surv(fi, di) ~ exp(x), survdat)
coxfit5 <- coxph(Surv(fi, di) ~ log(x), survdat)
coxfit6 <- coxph(Surv(fi, di) ~ I(1/x), survdat)  \# Note I(1/x)
coxfit7 <- coxph(Surv(fi, di) ~ sqrt(x), survdat)
# Calculating AIC for each of the 7 models
aic <-  c(-2*coxfit1$loglik[2] + (2 * 1), -2*coxfit2$loglik[2] + (2 * 1),
        -2*coxfit3$loglik[2] + (2 * 1), -2*coxfit4$loglik[2] + (2 * 1),
```

```
        -2*coxfit5$loglik[2] + (2 * 1), -2*coxfit6$loglik[2] + (2 * 1),
        -2*coxfit7$loglik[2] + (2 * 1))

# Calculating Likelihood Ratio Statistics(LRT) for each of the 9 models
lrt <- 2*c((coxfit1$loglik[2]-coxfit1$loglik[1]),(coxfit2$loglik[2]-
       coxfit2$loglik[1]),(coxfit3$loglik[2]-coxfit3$loglik[1]),
       (coxfit4$loglik[2]-coxfit4$loglik[1]),(coxfit5$loglik[2]-
       coxfit5$loglik[1]),(coxfit6$loglik[2]-coxfit6$loglik[1]),
       (coxfit7$loglik[2]-coxfit7$loglik[1]))
pval <- (1-pchisq(lrt,c(1,1,1,1,1,1,1)))
return(list(``aic"=aic,``pval"=pval,``lrt=lrt"))
```

A.4 Drawing the Histograms

```
# Codes used to draw the Histogram
histogram(~pvalmat, nint = 20, aspect = 1, ylim=c(0,15), xlim=c(0,1.0),
xlab = "P-values")
title("Histogram for 7,000 p-values, pooled from 7 models and 1,000
samples")
segments (0,0.346,1.0,0.346)
histogram(~pvalaic, nint = 20, aspect = 1, ylim=c(0,15), xlim=c(0,1.0),
xlab = "P-values")
title("Histogram for 1,000 p-values corresponding to best AIC models")
segments (0,0.346,1.0,0.346)
histogram(~pvalaicaw, nint = 20, aspect = 1, ylim=c(0,15), xlim=c(0,1.0),
xlab = "P-values")
title("Histogram for 1,000 p-values corresponding to Weighted models")
segments (0,0.346,1.0,0.346)
histogram(~pvalaicuw, nint = 20, aspect = 1, ylim=c(0,15), xlim=c(0,1.0),
xlab = "P-values")
title("Histogram for 1,000 p-values corresponding to Un-Weighted models")
segments (0,0.346,1.0,0.346)
```