Investigation of Over-fitting and Optimism in Prognostic Models

By

Matthew Richardson

A Thesis submitted to The University of Birmingham for the degree of Doctor of Philosophy

School of Health and Population Sciences

The University of Birmingham

March 2009

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

This work seeks to develop a high quality prognostic model for the CARE-HF data; see (Richardson *et al.* 2007). The CARE-HF trial was a major study into the effects of cardiac resynchronization. Cardiac resynchronization has been shown to reduce mortality in patients suffering heart failure due to electrical problems in the heart. The prognostic model presented in this work was motivated by the question as to which patient characteristics may modify the effect of cardiac resynchronization. This is a question of great importance to clinicians. Efforts are made to produce a high quality prognostic model in part through the application of methods to reduce the risk of over-fitting. One method discussed in this work is the strategy proposed by Frank Harrell Jr. The various aspects of Harrell's approach are discussed. An attempt is made to extend Harrell's strategy to frailty models. Key issues such as missing data and imputation, specification of the functional form of the model, and validation are examined in relation to the prognostic model for the CARE-HF data. Material is presented covering survival analysis, maximum likelihood methods, model selection criteria (AIC, BIC), specification of functional form (cubic splines and fractional polynomials) and validation methods (cross-validation, bootstrap methods). The concepts of over-fitting and optimism are examined. The author concludes that whilst Harrell's strategy is valuable it is still quite possible to produce models that are over-fitted. MDL (Minimum Description Length) is suggested as potentially useful methods by which statistical models can be obtained that have an in built resistance to over-fitting. The author also recommends that concepts such as over-fitting, optimism and model validation are introduced earlier in more elementary courses on statistical modelling.

Acknowledgements

The author should like to express his sincere thanks to the following people for their help and encouragement. Professor Nick Freemantle, Dr Melanie Calvert, Professor Jayne Parry, Mrs Emma Richardson and Mr Phillip Wilkes.

The author acknowledges his debt to Professor Nick Freemantle, Dr Melanie Calvert, Professor John G.F. Cleland and Prof Luigi Tavazzi for their efforts in collaborating on the development of the prognostic model for the CARE-HF data and their advice and assistance in writing the paper (Richardson *et al.* 2007) in which the model was first presented. The author wishes to thank the CARE-HF steering committee and Medtronic Inc.

The author wishes to thank Mr Domenico Pagano and Professor Nick Freemantle for granting their permission to the author to adapt a figure from their paper, see (Pagano *et al.* 2009) for inclusion in this work. The figure adapted from their paper appears as figure 4.1 in Chapter 4 of this work.

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 6

Chapter 7

Chapter 8

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 8

_Toc230078803

CHAPTER 1 INTRODUCTION

- Purpose of this work to develop a high quality prognostic model for the CARE-HF data
- A good prognostic model depends upon specifying an appropriate functional form
- Quality of a prognostic models depend upon important predictors not being omitted
- A poor prognostic model will result if depends there is a large amount of missing data
- Over-fitting leads to poor prognostic models
- Prognostic models should be validated

1.0.0 Introduction

This present work comprises in the main the development of a prognostic model for the CARE-HF data (Richardson *et al.* 2007). The CARE-HF trial is a landmark trial into the benefits of cardiac resynchronization therapy. Cardiac resynchronization therapy has been shown to significantly reduce mortality in patients suffering heart failure due to electrical abnormalities in the heart (Ellenbogen *et al.* 2005), (Cleland *et al.* 2001). In the next chapter I shall describe in further detail the background and development of this model. Briefly the model developed in (Richardson *et al.* 2007) aims to identify possible treatment modifiers of cardiac resynchronization therapy (Medtronic. 2009). It is of great importance that those patient characteristics which may modify the beneficial effects of cardiac resynchronisation are identified, i.e.

subgroups of patients are identified who may enjoy the most benefit from cardiac resynchronization therapy. My main aim is to produce a model that has been developed with the aim of minimising the risk of over-fitting and maximising its predictive power. This will entail amongst other techniques an application of an approach suggested by Frank Harrell Jr. I attempt later in this work to apply Harrell's approach in fitting a frailty model, an issue which Harrell does not address. I also seek to identify some of the limitations of Harrell's methods. My main objective is the development of a high quality prognostic model for what is an important real world application. The purpose of including some of the more theoretical material is to provide a framework in which I can understand issues that arise in developing a prognostic model.

1.1.0 Prognostic Models

I shall be concerned almost exclusively with prognostic models in this thesis. A prognostic model can be regarded as a tool by which a doctor can produce a prognosis for a patient. Prognosis from the Greek πρόγνωσις, can be defined as a doctor's prediction of how a patient's illness will develop and their chance of recovery. For example; given a patient's age, weight, blood pressure, a doctor could determine what if any beneficial effect a patient might experience if he or she were to receive a particular treatment or therapy. For general discussion of prognostic models the reader is directed toward Abu's paper (Abu & Lucas 2001). A prognostic model is a predictive tool; its purpose is to predict the level of increase in a beneficial effect, or the decrease in risk of some adverse event, for instance, death A prognostic model can assist a doctor in making clinical decisions, for example in trying to determine which patients might benefit from a particular treatment or therapy given that the treatment

is costly. The following papers provide excellent material on prognostic models (Wyatt & Altman 1995), (Moons *et al.* 2009) and (Royston *et al.* 2009).

The decision as to whether a patient will be given a particular treatment may well be based upon evidence obtained through the application of a prognostic model. Therefore, the importance of being able produce a reliable and accurate model is immediately seen. A method by which it is possible to assess the predictive accuracy of the prognostic model is also required. What steps can be taken in order to maximise the chances of producing a good model? These questions have led researchers to formulate a number of approaches to the modelling process with the aim of obtaining a parsimonious model that does not suffer from over-fitting and has good predictive accuracy.

1.2.0 Survival Analysis Background

If a new drug or treatment has been developed an important question is how effective is the drug or the treatment? Evidence for the efficacy of a drug or treatment is gathered by setting up a clinical trial. A simple situation might be as follows: A sample of patients suffering from some disease or illness is obtained. Patients from this sample are then randomly allocated to one of two groups. The first group is called the treatment group; patients allocated to this group receive the drug or treatment. The second group is called the control group, patients allocated to this group do not receive the drug or treatment, they may for instance be given a placebo. A researcher might then consider how many patients died in the treatment group compared to the control group (or more positively how many patients did not die). In the simple situation described above a researcher might use logistic regression to estimate the probability of death, the model might include variables such as patients age, sex. Also if a variable indicating to which group the patient belonged was included in the

model, and was subsequently found to be statistically significant then this may provide evidence for a treatment effect, i.e. probability of a patient dying is dependent upon whether or not they have received the treatment. In the example above the outcome is a binary one, dead or alive. It might well be that a patient's life is prolonged by taking the drug, but by how long? It might be a few months or it could be 20 years. In many clinical trials the question of the efficacy of a drug or treatment is addressed in terms of the time to event, i.e. how long until a patient dies or experiences the event of interest. In this case survival analysis is the appropriate method. A few words should be said on the matter of randomisation. One of the principal reasons for adopting randomisation when developing a prognostic model is to avoid biased estimates of treatment effects. In attempting to develop a prognostic model it is important that the treatment and control groups are balanced in terms of the distribution of variables that may be strong predictors of the outcome. Randomisation also reduces the risk of obtaining biased estimates of the treatment effect due to missing or unknown variables. It should be borne in mind that randomisation does not guarantee that estimates of treatment effects will be unbiased in all situations (Gail *et al.* 1984). I should like to point out however that randomisation can be considered as a controversial topic (Royall 1991). However R.A Fisher (Fisher 1966) argues that if we assume that a real treatment effect is absent, then the result from any experiment is due to chance alone. Fisher (Fisher 1966) provides a very clear argument to support of randomisation. I shall now review some of the fundamental ideas in Survival analysis, what follows is a standard derivation of the basic results. I make no claim whatsoever to have developed anything new. These are well known results attributable to others. Similar derivation may be found in any number of statistics textbooks, see Dobson's textbook (Dobson 2002) for example.

Let $Y > 0$ be the survival time or time to failure (or some event).

Then $P(Y < y) = F(y) = \int_0^y f(u)du$, also $P(Y \geq y) = 1 - F(y) = S(y)$.

$S(y)$ is the survivor function. The probability per unit time of failure occurring in the interval $(y, y + \delta y)$ given survival up-to time y is given by the hazard function $h(y)$.

$$h(y) = Lim_{\delta y \to 0} \frac{P(y \leq Y < y + \delta y \mid Y \geq y)}{\delta y}$$

Now we have $P(y \leq Y < y + \delta y) = \int_y^{y+\delta y} f(u)du = F(y + \delta y) - F(y)$.

Note that $P(y \leq Y < y + \delta y) = P([y \leq Y < y + \delta y] \,\&\, [Y \geq y])$.

The conditional probability $P(y \leq Y < y + \delta y \mid Y \geq y)$ can be expressed as

$$\frac{P([y \leq Y < y + \delta y] \,\&\, [Y \geq y])}{P(Y \geq y)} = \frac{F(y + \delta y) - F(y)}{S(y)} \quad .$$

This gives

$$h(y) = Lim_{\delta y \to 0} \frac{P(y \leq Y < y + \delta y \mid Y \geq y)}{\delta y} = Lim_{\delta y \to 0} \frac{F(y + \delta y) - F(y)}{S(y)\delta y} = \frac{f(y)}{S(y)}.$$

Differentiating $\log_e S(y)$ we get $\frac{d}{dy} \log_e S(y) = \frac{-f(y)}{S(y)} = -h(y)$.

It follows that $-\int_0^y h(u)du = \log_e S(y)$, giving $e^{-\int_o^y h(u)du} = S(y)$.

$\int_0^y h(u)du = H(y)$ is the cumulative hazard function.

The exponential distribution is a sensible candidate for the distribution of Y.

So $f(y) = \theta e^{-\theta y}, y \geq 0, \theta > 0, E(Y) = \frac{1}{\theta}, \text{var}(Y) = \frac{1}{\theta^2}$.

Then $F(y) = \int_0^y \theta e^{-\theta u} du = 1 - e^{-\theta y}$, $1 - F(y) = S(y) = e^{-\theta y}$.

Using $\frac{d}{dy} \log_e S(y) = -h(y)$ we get $h(y) = \theta$, using $\int_0^y h(u)du = H(y)$ to

obtain $H(y) = \theta y$.

For the exponential distribution consider Y as depending on some variables $x_1, x_2, ...$

Then $E(Y) = \tilde{\beta}\tilde{x}$ say, but $\theta > 0$, so we use $E(Y) = e^{-\tilde{\beta}\tilde{x}}$, i.e. $\theta = e^{\tilde{\beta}\tilde{x}}$.

Writing $h(y) = e\tilde{\beta}\tilde{x} = e^{\sum \beta_i x_i}$, let x_j be an indicator variable, $x_j = 0$ denote absence,

$x_j = 1$ denote presence of some exposure.

$\frac{h_1(y)}{h_0(y)} = e^{\beta_j}$ is the hazard ratio.

In general models of the form $h_1(y) = h_0(y)e^{\tilde{\beta}\tilde{x}}$ are known as proportional hazards

models. For the proportional hazards model

$$H_1(y) = \int_0^y h_1(u)du = \int_0^y h_0(u)e^{\tilde{\beta}\tilde{x}} du = H_0(y)e^{\tilde{\beta}\tilde{x}}$$

Taking logs gives

$\log_e H_1(y) = \log_e H_0(y) + \tilde{\beta}\tilde{x} = \log_e H_0(y) + \sum \beta_i x_i$.

Considering the indicator variable x_j we can write $\log_e H_1(y) = \log_e H_0(y) + \beta_j$, the natural logarithms of the cumulative hazard functions differ by the constant β_j. For proportional hazards a plot of $\log_e(y), \log_e H_1(y)$ and $\log_e(y), \log_e H_1(y)$ on the same set of axis should show parallel lines.

Certain subject may survive beyond the duration of the study, these cases are said to be censored. With censored cases we cannot have full knowledge regarding survival time; all we can say is that the subject survived up to the end of the study. The subject may well experience failure the day after the study ended, or 2 months later. Let $y_1, y_2, ..., y_k$ be the survival time for the uncensored cases, and let $y_{k+1}, ..., y_n$ be the survival times for the censored cases. Further let $\zeta_i = 1$ for the uncensored cases and $\zeta_i = 0$ for the censored cases, then the likelihood function L is given by

$L = \prod_i^n f(y_i)^{\zeta} \; S(y_i)^{1-\zeta_i}$.

The log likelihood $\log_e(L)$ is given by

$$\log_e(L) = \sum_i^n \zeta_i \log_e(f(y_i)) + (1-\zeta_i)\log_e(S(y_i)) .$$

Both L and $\ln(L)$ depend on the parameters of the distribution y and on $\widetilde{\beta}\widetilde{x}$, this is a parametric model. $\ln(L)$ can be maximised using the Newton-Raphson method. In the Cox Proportional Hazards model $f(y)$ and $S(y)$ are not completely defined, in this case the model is described as being non-parametric (distribution of y not specified)

1.3.0 The Modelling Process

Harrell et al. (Harrell *et al.* 1996) identify the following as potential problems in the modelling process:

- Violation of Assumptions
- Omission of Important Predictors
- Missing Data / Incorrect Imputation
- Over-fitting

Each of the above may lead to an ill-fitting prognostic model; predictions based on such a model will not be reliable. When attempting to fit any mathematical or statistical model the researcher is often compelled to make a number of simplifications and assumptions. Real world situations are often too complex to model without such simplifications and assumptions. In fitting a prognostic model three basic assumptions shall be made; the first is a distributional assumption, the second an assumption regarding functional form and the third an assumption about additivity.The prognostic models that I shall consider in this thesis are based on the Cox Proportional Hazards model (Cox 1959), (Cox 1964), (Cox 1972), of course prognostic models can be developed for other forms of Generalised Linear Models (GLM) see (Nelder & Wedderburn 2009), (Baker & Nelder 1978) and (Dobson 2002). In a linear model $E(Y) = \widetilde{\beta x}$, the GLM extends the linear model to situations where the relationship between $E(Y)$ and $\widetilde{\beta x}$ is not linear, this is achieved through the link

function $f_L()$, so that $f_L(E(Y)) = \widetilde{\beta}\widetilde{x}$. Prognostic models could for example be based on other regression models, e.g. logistic.

1.4.0 Violation of Assumptions

Although this thesis is concerned with over-fitting and optimism it is considerable importance that basic assumptions are examined as to their validity. For example with the Cox model is the proportional hazards assumption valid? Are assumptions about the function form of the model appropriate? Once a model has been obtained is it clinically plausible?

1.4.1 Proportional Hazards

Under the proportional hazards assumption the hazard ratio $\frac{h_1(y \mid \widetilde{x}_1)}{h_0(y \mid \widetilde{x}_2)}$ is a constant over time, the effect of a covariate does not vary over time. The Cox proportional hazards model makes no assumptions about the form of $h(y)$ and is described as semi-parametric, this is an advantage of the Cox model in that it is possible to avoid specifying an inappropriate form for $h(y)$.Estimates for the Cox model are obtained through partial likelihood (Cox 1972). If $h(y)$ is specified for example $h(y) = \exp(u + v\log_e(y))$ then certain assumptions about the distribution of the survival time Y have to be met, in this case Y follows the Weibull distribution.
It may not be reasonable to assume proportional hazards. If the proportional hazards assumption is violated, it is possible to include a time dependent variable in the model; however inclusion of a time dependent variable leads to difficulties in assessing the validity of the model. Another strategy for dealing with non proportional

hazards is to stratify the model based on the variable for with the proportional hazards assumption is violated.

1.4.2 Functional Form

Specifying an appropriate functional form for the model is important. The assumption of a simple relationship between Y and X such as $Y = X$ may not be appropriate. There may be a more complex relationship between Y and X. In this situation it is required to transform X, examples of typical transformations are $\log_e(X)$, $\sqrt{X}$. However the crucial point is that the model is linear in the parameters. The models $Y = \beta_0 + \beta_1 X_1 + error$ and $Y = \beta_0 + \beta_1 X_1^{\,2} + error$ are both linear models i.e. the right hand side in both cases is a linear combination of the parameters β_0 and β_1. In recent years a great deal of work has been carried out in the study of cubic splines and their application to statistical models. There are instances when the fit of a model can be improved by using cubic splines in the specification of the functional form. The following authors provide very useful material on the use of cubic splines in statistical modelling, (Wegman & Wright 1983), (Smith 1979), (Poirier 1979), (Royston 2000) and (Herndon & Harrell 1990). Another extremely interesting approach to transformations is that of the Fractional Polynomial (Royston & Altman 1994), (Royston Patrick *et al.* 1999) and (Royston & Sauerbrei 2004). The reader is encouraged to read Royston and Altman's paper (Royston & Altman 1994), further useful material can be found in (Royston *et al.* 1999), (Royston & Sauerbrei 2004) and software for fitting fractional polynomials is documented in (Meier-Hirmer *et al.* 2003). Both cubic splines and fractional polynomials will be covered in greater detail in Chapter 4.

1.4.3 Additivity

For the Cox proportional hazards model the relationship

$\log_e\left(\frac{h_1(y)}{h_0(y)}\right) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$ could be treated as a multiple regression model then the change in the expected value of $\log_e\left(\frac{h_1(y)}{h_0(y)}\right)$ due to a unit change in the independent variable X_i (holding all other $X's$ constant) is the same regardless of the values (levels) of the other independent variables, if this is assumed the effects are said to be additive. In simple and multiple linear regression it is the additivity assumption allows the interpretation of β_i as the change in $E(Y)$ due to a unit increase in X_i given that all the other $X's$ are held constant without specifying at what value the other $X's$ are held constant. The assumption of additivity is perhaps too restrictive, it may well be that changes in the expected value of $\log_e\left(\frac{h_1(y)}{h_0(y)}\right)$ for a unit change in X_i are dependent on the values of one or several of the other independent variables. Relaxing the assumption of additivity requires that interaction terms be introduced into the model.

If it is found that the additivity assumption is violated, then clinically/biologically meaningful interaction terms should be included in the model. In the simple model $Y = \beta_1 X_1 + \beta_1 X_2 + \beta_3 X_1 * X_2 + \beta_0$, let X_1 be the patients age in years, and X_2 indicate whether the patient has received treatment or not, then the term $X_1 * X_2$, represents an interaction term, the interaction term describes how age modifies the effect of treatment.

1.5.0 Omission of Important Predictors and Missing Data

1.5.1 Omission of Important Predictors

The omission of important predictors can lead to an inaccurate model in the sense that estimates of treatment effect will be biased. It may be that some important predictor of outcome is as yet unidentified, or it is a known predictor and has been omitted for some reason. Randomisation offers a way of reducing the risk of biased estimates for the treatment effect when important predictors have been omitted for whatever reason.

1.5.2 Missing Data

Missing data will have a bearing on the final model, distorted estimates of predictor variables may result from missing data. A variable that appears not to be statistically significant due to a high level of missing data, may in fact be of considerable predictive value. Missing data may categorised as being missing completely at random, missing at random, and missing not at random.

It is important that the missing data mechanism is identified. For an introduction to some of the terminology used in connection with missing data see the website operated by LSHT (London School of Hygiene and Tropical Medicine 2008). Missing data and imputation will be discussed in Chapter 7. For the present, suffice it to say that once the nature of the missing data has been established steps can be taken to deal with this problem, i.e. the missing data is imputed. It is important that the correct imputation method is applied. Imputation is a complex problem, for a detailed treatment of developing prognostic models when missing data is present see (Marshall 2007).

1.6.0 Over-fitting and Optimism

Over-fitting may be described in the following way. In dealing with a binary outcome, for example dead or alive, it may be that interest is focused on predicting deaths, the ratio of deaths/predictor degrees of freedom can be used to gauge the level of what is known as over-fitting. If the number of events of interest is small and a large number of independent variables are included it is likely that the model will be over-fitted. It will be found that independent variables are included in the model (deemed statistically significant) due to their being 'locally important predictors'. On validating the model it may be found that these independent variables are not significant. In over-fitting a model, noise and localized features in the data attain a spurious statistical significance and lead to biased model. Considering predictive accuracy when over-fitting is present, this means that the predictive accuracy of the model when validated on an external dataset will be seen to deteriorate. The predictive accuracy of the model using the data on which it was developed may be quite good; yet when the model is applied to a new (but similar) data set it is found that the predictive accuracy is poor in comparison, this is known as optimism or statistical optimism.

1.7.0 Data Reduction and Shrinkage

Data reduction can be described as a means of reducing in the number of independent variables that might be included during the modelling process (reduction of the dimensions of the data). If an attempt is made to fit a model with 70 variables to a data set of 50 patients, then the model will be severely over-fitted. By employing data reduction it may be possible to reduce the risk of over-fitting, a classical data reduction technique is principal components analysis, see Sharma (Sharma 1995). Empirical rules have been arrived at which can be applied to determine if data

reduction should be used. One such rule for the Cox proportional hazards model is based on the ratio $\frac{N_E}{p}$ (events per variable), where N_E is the number of uncensored events, and p is the predictor degrees of freedom, p can be thought of as the number of independent variables included in the model. If $\frac{N_E}{p} < 10$ there is some risk of over-fitting, consequently we should look to performing data reduction, i.e. reduce the number of independent variables that are included in the model, (Peduzzi *et al.* 1996) provides background on events per variable rules. In a good model a linear relationship should be observed between the observed (i.e. new data) and predicted (i.e. predictions made using the original data) values, i.e. $Y = \hat{Y}$ (the line has a slope of 45^o and passes through the origin), departure from a slope of 45^o indicates that over-fitting has occurred. Over fitting is not the only cause of a departure from the 45^o slope, for instance if assumptions relating to the error term in the model have been violated a departure from will be observed, for example term the error does not have a constant variance, the error terms are not independent. This departure from the 45^o slope due to over-fitting is known as shrinkage, a measure of the shrinkage gives a measure of over-fitting. Van Houwelingen and le Cessie (Van Houwelingen & le Cessie 1990) have developed a heuristic estimator of shrinkage $\hat{\gamma} = \frac{\chi^2 - p}{\chi^2}$, here χ^2 is the total model log likelihood ratio statistic used in testing for associations between X and Y, it can be seen that as p the predictor degrees of freedom decreases so does $\hat{\gamma}$.

Use of the entire data set in developing the model allows for the extraction of maximum information, as Harrell (Harrell *et al.* 1996) points out “data are too

precious to waste". Outliers or highly influential observations offer some clues about possible over-fitting. If for some X there exist one or two extreme values, it may be that X appears as a significant predictor, these extreme values can lead to the selection of spurious predictors, resulting in a model that has been over-fitted, the model is not general. After validation X may be found to be not significant, the extreme values in a particular data set 'drove' the modelling process.

1.8.0 Validation

Once a model has been obtained it should be validated using a new data set. I said earlier that the predictive accuracy of the model for the data on which it was developed may be quite good, however this is not sufficient to claim that the model is generally good (Altman & Royston 2000). The model must be validated using new data, even if it is found that for the original data the model performs well. A model that performs well on the original data set is not guaranteed to perform well when applied to a new but similar set of patients. This point may appear to have been laboured somewhat, but it is crucial when fitting a model we have in mind the idea of generalisabilty. Bootstrapping, validation, calibration and discrimination (component parts of predictive accuracy) will be discussed in Chapter 6. Over the years many other statisticians have been engaged in research into the problem of producing reliable models that do not suffer from gross over-fitting and possess good predictive accuracy. I shall now briefly outline some of this work

1.9.0 Harrell et al.'s Approach

Harrell et al. (Harrell *et al.* 1996) have devised a systematic approach to fitting a prognostic model which may be summarised as follows:

- Obtain an accurate and large sample of data.
- Formulate a sharp hypothesis.
- Discard observations with missing Y, provided Y missing at random.
- For missing X investigate factors related to missingness, if the number of observations that would be excluded is small or variable that is missing is unimportant, then exclude observations with missing values. Otherwise impute missing X.
- If the number of variables included in the model is large in comparison to the number of events of interest, use data reduction.
- Use the entire sample to develop the model.
- Check linearity assumptions and perform transformations on Xs if required.
- Check additivity assumptions, include clinically motivated interaction terms.
- Check for outliers or influential observations.
- Check distributional assumptions, for Cox Proportional Hazards model, proportional hazards assumption, if violated include time dependent variable.
- Perform backward stepdown variable selection.
- Variables obtained from stepdown procedure form the final model.
- Validate model using the bootstrap.
- If using stepwise variable selection, supply a Table showing how important predictors vary over the bootstrap samples.
- Estimate shrinkage.

1.10.0 Ambler, Brady and Royston's work

Ambler, Brady and Royston have investigated methods for estimating and simplifying

full models (Ambler *et al.* 2002). In (Ambler *et al.* 2002) the authors aim to produce simplified models that retain their prognostic power. Based on simulation studies using two different data sets, Ambler *et al* state that the results of model simplification based on the stepdown variable selection, using maximum likelihood and penalised maximum likelihood depended upon whether or not all the independent variables (predictors) were influential. Harrell advocates limited variable selection based on the stepdown method. The stepdown method makes use of the idea of a prognostic index, if $X_1, X_2 ... X_p$ are independent variables then the prognostic index is a linear combination of $X_1, X_2 ... X_p$, i.e. $a_1 X_1 + a_2 X_2 + ... + a_p X_p$. Regression of the linear combination on the independent variables results in a perfect fit, $R^2 = 1$, if any of the independent variables are omitted, then R^2 will decrease. A simplified prognostic index is defined as the linear combination formed by removing the X_m which causes the smallest decrease in R^2, this process is carried out until further removals of X_m would result in $R^2 < \alpha$, where α is a predefined value for R^2. Ambler *et al* suggest the Akaike Information Criteria (AIC) provides a good way of selecting a simplified model, the background to the AIC will be presented in Chapter 5.). In my review of Harrell's approach I referred to events per variable, and how by reducing p it is possible to avoid over-fitting. Models produced using a criterion such as $\frac{N_E}{p} < 10$ are known as full models, the complexity or size of the model is determined by the number of events of interest in the data. Ambler *et al* make a very important point; full models are liable to be very complex when we have data containing a large number of observations and a large number of possible predictor variables. Large and complex models have attached to them financial and practical drawbacks. This may be seen as a drawback to Harrell's approach.

1.11.0 Van Houwelingen's work

Hans C. van Houwelingen describes methods for determining the predictive accuracy of prognostic survival models; see (van Houwelingen 2000). Van Houwelingen develops what he calls validation by calibration; he illustrates this method by the following example using simple linear regression:

- Plot Y against $\hat{Y} = X'\beta_{\mathrm{model}}$ for new data.
- If $Y = \hat{Y}$ appears to hold (points lie on 45^o line through (0,0)), then model is valid.
- If $Y = \hat{Y}$ does not appear to hold (points do not lie on 45^o line through
- (0,0)), correct model by calibration.
- Fit $Y = \alpha + \beta\hat{Y} + e$, then $Y_{cal} = \hat{\alpha} + \hat{\beta}\hat{Y}$ is the calibrated model.

Van Houwelingen explains that his strategy is to compare a particular model with the new (validation) data set and not a new model obtained from the validation data set. From a theoretical perspective his method is appealing in its simple and clean approach. Van Houwelingen proposes a method whereby the Cox proportional hazards and the non-proportional hazards model may be calibrated (van Houwelingen 2000).

1.12.0 Extension of Cox Proportional Hazards Model

Initially I shall develop a prognostic survival model using Cox Proportional Hazards model, in a later Chapter I will look at how Cox Proportional Hazards model can be extended to deal with heterogeneous data through the use of frailty (Vaupel *et al.* 1979). The frailty model is an interesting advance in modeling. Harrell *et al's* approach as far as this author is aware does not address frailty. In Chapter 8 I attempt to fit a frailty model using Harrell's approach.

1.13.0 Current Modelling Strategies

Are issues such as over-fitting and specification of functional form for a prognostic model routinely addressed? Has the approach suggested by Harrell been widely adopted? I have carried out an informal survey of three journals; BMJ, JAMA and Circulation. Papers were selected from these journals based on the key words such as prognostic, survival, Cox model, and risk score. It appears that over-fitting is not routinely addressed. Where it might be appropriate specification of functional form using cubic splines or fractional polynomials is not widely adopted practice.

It is hoped that the reader may glean some practical guidance on how to employ Harrell *et al's* approach, and perhaps become aware of some of the difficulties that can arise. If by reading this thesis the reader who may not be a statistical expert, acquires a better understanding of the important issues surrounding the development of a prognostic model then I would have accomplished a main objective. That everyone should take steps to ensure that the chances of over-fitting a model are minimised is of course highly desirable. However it may be worth trying to consider some of the natural and inherent limitations to the statistical method. There is no correct model, every model is an approximation; to quote G.E.P Box, "Essentially, all models are wrong, but some are useful." (Box & Draper 1987). The whole question of generalizabilty is a complex one. Should we expect to achieve more general results in the physical sciences? In the next Chapter I shall consider the development of a prognostic survival model for the Cardiac Resynchronization Therapy in Heart Failure (CARE-HF) data set.

CHAPTER 2 THE CARE-HF STUDY

- Cardiac resynchronization therapy significantly reduces mortality in patients with heart failure
- What patient characteristics may modify the effects of cardiac resynchronization therapy?
- To investigate treatment modifiers a prognostic model is developed
- Ischaemic aetiology, more severe MR, and increased NT-pro-BNP were all independent predictors of an increased risk of death or unplanned cardiovascular hospitalization irrespective of randomised treatment (CRT)
- Systolic blood pressure and Interventricular mechanical delay are identified as treatment modifiers

2.0.0 Introduction

In this work a prognostic model was fitted to data obtained from the Cardiac Resynchronisation in Heart Failure Trial (CARE-HF). CARE-HF is one of the largest randomized studies of cardiac resynchronization therapy (CRT), has a longer duration of follow-up than any other, and has a robust primary clinical endpoint (Richardson *et al.* 2007). These attributes make it a valuable resource for the investigation of those factors that predict the likelihood that a patient will or will not respond to cardiac resynchronisation therapy (CRT). Clinicians view CRT in the context of those patients who will derive benefit from CRT (responders) and those who will not (non responders). If a patient is in receipt of CRT what characteristics of that individual may determine the likelihood of them receiving benefit from the treatment? This leads

us (Richardson *et al.* 2007) to consider treatment modifiers, i.e. those patient attributes that modify the effect of CRT. CRT is a treatment that aims to restore and improve cardiac function in patients who suffer electrical conduction problems in the heart as a result of heart failure (Medtronic 2009). Heart failure is a common and serious condition with a complex and varied pathophysiology (Cleland *et al.* 1999). A substantial minority of patients with heart failure due to left ventricular (LV) systolic dysfunction have prolonged QRS, QRS represents ventricular depolarisation and amongst these patients there is a high prevalence of cardiac dyssynchrony, which leads to a decline in cardiac efficiency through diverse mechanisms, see (Xiao *et al.* 1993), (Daubert *et al.* 1999) and (Auricchio *et al.* 1999). For patients with heart failure due to cardiac dyssynchrony who have persistent moderate or severe symptoms despite standard pharmacological therapy, CRT improves cardiac function leading to an improvement in well-being and a reduction in morbidity and mortality, see (Abraham *et al.* 2002),(Bristow *et al.* 2004),(Cleland *et al.* 2005) and (Freemantle *et al.* 2006).

CRT is delivered by means of a physical device akin to a pacemaker, see (Medtronic 2009). The aim of this analysis was to evaluate the relationship between prospectively defined clinical, echocardiographic and neurohormonal variables, collected at baseline during the CARE-HF trial, on overall outcome in all patients and on the response to CRT.

The prognostic model presented in this work, is that developed by Richardson, Freemantle, Calvert, Cleland and Tavazzi (Richardson *et al.* 2007) based on Individual patient data collected during the CARE-HF trial. The design and results of the CARE-HF study have been reported previously (Cleland *et al.* 2005), (Cleland *et*

al. 2001). In brief, the CARE-HF trial enrolled 813 patients recruited from 82 centres across Europe. Eligible patients were at least 18 years of age, had evidence of heart failure for at least 6 weeks, and were in New York Heart Association class (NYHA) III or IV despite receipt of standard pharmacologic therapy, with a LV ejection fraction (EF) of < 35%, a LV end-diastolic dimension of ≥ 30 mm (indexed to height), and a QRS interval of > 120 ms on the electrocardiogram. Patients with a QRS interval of 120–149 ms were required to meet two of three additional criteria for dyssynchrony: an aortic pre-ejection delay of more than 140 ms, an interventricular mechanical delay (IVMD) of > 40 ms, or delayed activation of the posterolateral LV wall. The IVMD was calculated as the time difference between the onset of forward flow in the LV (APET) and RV (PPET) outflow tracts: IVMD =APET – PPET (Ghio *et al.* 2006). A total of 409 patients were randomized to CRT and medical therapy, whereas 404 received medical therapy alone (Richardson *et al.* 2007). The primary outcome was the time to death from any cause or an unplanned hospitalization for a major cardiovascular event. Patients were followed up for a mean of 29.4 months.

2.1.0 Developing the Prognostic Model

A number of potentially important clinical, echocardiographic, and neurohormonal variables collected at baseline were specified a priori for evaluation in a prognostic model. These were mitral regurgitation (MR), end-systolic volume index, aetiology (ischaemic and non-ischaemic disease), EF, use of beta-blockers, age, QRS interval (QRS), supine systolic blood pressure (SBP), glomerular filtration rate, N-terminal pro-brain natriuretic peptide, as determined by Roche Assay (NT-pro-BNP), and IVMD ,see (Talwar *et al.* 1999), (Pitzalis *et al.*. 2005), (Doust *et al.* 2005). MR was defined as area of colour flow Doppler regurgitant jet divided by area of left atrium in systole, both in square centimetres. The primary composite outcome was time to death from any cause, or an unplanned hospitalization for a major cardiovascular event. Cox Proportional Hazards models were fitted to identify predictors of risk of death from any cause or an unplanned hospitalization for a major cardiovascular event (main effects) and to identify any predictors modified by cardiac resynchronization (Hosmer & Lemeshow 1992) and (Lee 1992) ,the SAS code for producing theses models is to be found in Appendix 1.0.0. The modelling strategy was based upon the approach suggested by Harrell *et al* 1996, see Chapter 1 for an introductory discussion of Harrell's approach. In order to evaluate whether any of the variables had a non-linear

relationship with outcome, transformations of each variable using the natural logarithm and cubic spline were assessed (Herndon & Harrell 1990), (Wegman & Wright 1983), (Poirier 1979), (Smith 1979) and (Royston 2000) see Chapter 4 for a further discussion of cubic splines, SAS code used for fitting cubic splines is to be found in Appendix 1.0.0. The Akaike Information Criterion (AIC) was used to determine the most appropriate transformation (Akaike 1974), see Chapter 5 for a more detailed discussion of the AIC. The validity of any transformations was further assessed by examining plots of the cumulative Martingale residuals versus the transformed variable (Verweij *et al.* 1998), (Therneau & Grambsch 1990). The proportional hazards assumption was also assessed. Statistically significant variables identified from univariate analyses (Table 2.2).

All analyses were performed in SAS v 9.1 using the PHREG procedure and the RCS macro (Heinzl & Kaider 2006). The RCS macro was used to fit cubic splines with four knots, Herndon and Harrell (Herndon & Harrell 1990) suggest based on empirical studies, that 4 knots are sufficient to model most data, this point will be considered further in Chapter 4. For the continuous variables, with the knot positions specified PHREG was then used to generate a model from which it was possible to determine whether the cubic spline was an appropriate transformation for the particular variable concerned. All analyses were undertaken according to the intention to treat principle, i.e. the effect of a treatment is assessed based on the planned treatment rather than the actual treatment (ICH E9. 1999). In a clinical trial use of the intention to treat principle allows for an unbiased estimate of the effect of a treatment

in situations where a number of patients may not adhere to the treatment programme. Alternative approaches are to exclude those patients who do not adhere or to include them in the group the treatment group, this approach leads to a biased estimate of the treatment effect (Montori & Guyatt 2001). To validate the final model two further steps were taken. First, a bootstrap revalidation process was used to estimate the degree of over-fitting from the model fitting process (Harrell *et al.* 1996). The design library in the statistical package R was used to undertake this validation (Design Library Harrell Frank E. 2009a). Second, multiple imputation using the SAS procedures MI (SAS Institute. 2009), and MIANALYSE were employed to examine the effect of missing data on the final model. In (Richardson *et al.* 2007) it must be stressed that the authors were concerned with identifying possible treatment modifiers i.e. interactions with CRT. The approach to identifying possible treatment modifiers presented in (Richardson *et al.* 2007) and this thesis are open to question and criticism. It can be argued that if there is a genuine interaction between CRT and another independent variable then this interaction will be identified using the conventional approach of first fitting main effects and then going on to fit interaction terms. The approach to identifying interaction terms adopted in this thesis had been employed in previous work and was suggested to myself as a way of dealing with fact that treatment modifiers where the primary concern as opposed to main effects. I do not claim that this approach is the right way. Those variables identified to be significantly ($P< 0.05$) associated with the primary composite outcome (time to death from any cause, or an unplanned hospitalization for a major cardiovascular event) were entered in a multivariable Cox Proportional Hazards model using a forward stepwise selection to obtain the final model (Table 2.3). The entry criteria for the forward selection procedure was 0.05, meaning a variable has to be significant at the

0.05 level before it can enter the model. When using a forward selection method I start by fitting the Cox models $h(y) = \exp(\beta_0 + \beta_i X_i)$, where *i*= (*1,2…,m), m* is the number of independent variables, i.e. I have the models

$h(y) = \exp(\beta_0 + \beta_1 X_1), h(y) = \exp(\beta_0 + \beta_2 X_2), \ldots, h(y) = \exp(\beta_0 + \beta_m X_m)$.

For each of these models once the p-value *p* for X_i was determined, I can identify candidate variables for inclusion in the model by considering all X_i where $p < \alpha$, α being a prescribed significance level. If there are several X_i that satisfy $p < \alpha$, then I select X_k, where X_k is the X_i with the smallest p-value from amongst the candidate X_i s. We then fit the models $h(y) = \exp(\beta_0 + \beta_k X_k + \beta_i X_i)$, $i \neq k$. From these models the X_i that has the smallest p-value (denoted X_l) is included i.e. I now have $h(y) = \exp(\beta_0 + \beta_k X_k + \beta_l X_l + \beta_i X_i) \quad i \neq k, i \neq l$. This process is repeated until there are no independent variables left. In the forward selection method a variable will remain in the model no matter what new variables are included. In the forward stepwise selection procedure the forward selection procedure described above is refined in the following way. The p-value *p* of each independent variable that is already included in the model is examined at each step. If *p* is greater α , then X_k is removed from the model. Also, if X_k has been removed previously from the model it may re-enter if *p* is less than α, but it may re-enter only once, it cannot enter more than twice. In forward stepwise selection I start by fitting the models $h(y) = \exp(\beta_0 + \beta_1 X_1), h(y) = \exp(\beta_0 + \beta_2 X_2), \ldots, h(y) = \exp(\beta_0 + \beta_m X_m)$ these models will under-fit the data. Harrell suggests that the 'limited' backward stepwise selection be employed; it is claimed that this method has advantages over the

forward stepwise selection, a comprehensive discussion on backward methods can be found in (Mantel 1970). Beale (Beale 1970) offers some interesting criticism of Mantel's arguments. In backward stepwise selection I start with the full model, which it could be argued is a better starting point. The choice between backward or forward stepwise selection is in my view a matter for the individual researcher, it would be misleading to dismiss forward selection without considering the fact that all automatic variable selection methods, including backward methods can be criticised as producing suspect models. Ira Berstein has described Stepwise methods as "data driven variable selection schemes", (Ulrich 1997), Harrell although suggesting that a researcher perform stepwise selection (Harrell *et al.* 1996) points out that stepwise methods do not tackle over-fitting, and recommends that variables are retained in the model irrespective of their p-values, as this leads to a model with better discriminatory power compared to a model produced solely on the basis of the stepwise selection method. This appears strange, variables that are not statistically significant and might be regarded as being redundant are important in terms of the discriminatory power of the model (they may be clinically significant). Forward stepwise selection is useful in situations where I might wish to fit a large number of interactions. Which selection method is best? A definite answer to this question does not appear to exist. All variable selection procedures posses some defect, and so whichever method a researcher adopts he or she must carefully examine the final model and perform some type of validation.

2.2.0 Results

The baseline characteristics of patients from the CARE-HF trial are shown in Table 2.1.

	Control			**Treatment**		
	n	**median**	**(IQR)**	**n**	**median**	**(IQR)**
Age (years)	403	66	(59–72)	409	67	(60–73)
Aetiology (ischaemic Y/N)	Y=153 N=250			Y=186 N=223		
Systolic blood pressure (mmHg)	399	110	(100–125)	404	110	(100–125)
Glomerular filtration rate (mL/min/1.73m^2)	372	61	(46–73)	367	60	(46–73)
N-terminal pro-brain natriuretic peptide (pg/ml)	370	1806	(719–3949)	362	1920	(744-4288)
Use of beta-blockers (Y/N)	Y=288 N=116			Y=298 N=111		
QRS width (ms)	394	160	(152–180)	401	160	(152–180)
Interventricular mechanical delay (ms)	370	50	(30–66)	365	49	(32–67)
End-systolic volume index (mL/m^2)	376	117	(94–147)	356	121	(92–151)
Ejection fraction (≤ 35%)	378	25	(22–29)	367	25	(21–29)
Mitral regurgitation	303	23	(11–34)	302	21	(12–33)

Table 2.1 Baseline characteristics of patients total number in study N=813

Notes for Table 2.1 IQR (interquartile range). Mitral regurgitation defined as area of colour flow Doppler regurgitant jet divided by area of left atrium in systole, both in square centimeters.

These data are consistent with the patients having, on average, moderate to severe LV systolic dysfunction, dilatation and dyssynchrony with a low arterial pressure, and renal dysfunction. About 40% of patients had ischaemic heart failure due to ischaemia. Univariate analyses were used to identify those variables that were significant predictors of outcome (time to death from any cause, or an unplanned hospitalization for a major cardiovascular event), irrespective of treatment allocation,

and those variables shown to predict response to CRT (indicated by the CRT * variable interaction term) (Table 2.2).

	n	**Hazard ratio**	**95% CI**	**P-value**
Mitral regurgitation [a]	605	2.14	1.68–2.71	0.0001
CRT		1.85	0.59–5.08	0.2938
CRT * Mitral regurgitation [a]		0.72	0.50–1.02	0.0670
Interventricular mechanical delay (ms)	735	0.99	0.99–1.00	0.0028
CRT		0.92	0.62–1.36	0.6784
CRT * Interventricular mechanical delay (ms)		0.99	0.99–1.00	0.0473
End-systolic volume index (mL/m2) [a]	732	1.52	1.08–2.14	0.0175
CRT		0.62	0.04–9.88	0.7354
CRT * End-systolic volume index (mL/m2) [a]		1.00	0.56–1.77	0.9978
Glomeruler filtration rate (ml/min/1.73 m2)	739	0.99	0.98–0.99	0.0005
CRT		0.74	0.38–1.48	0.3964
CRT * Glomeruler filtration rate (ml/min/1.73 m2)		1.00	0.99–1.01	0.5811
Systolic blood pressure (mmHg)	803	0.99	0.98–1.00	0.0011
CRT		0.14	0.03–0.63	0.0097
CRT * Systolic blood pressure (mmHg)		1.01	1.00–1.03	0.0491
Ejection fraction (%)[a]	745	0.38	0.22–0.66	0.0006
CRT		0.38	0.02–5.44	0.4298
CRT * Ejection fraction (%)[a]		1.24	0.51–3.03	0.6341
N-terminal pro-brain natriuretic peptide (pg/mL) [a]	732	1.47	1.31–1.66	0.0001
CRT		0.33	0.08–1.37	0.1275
CRT * N-terminal pro-brain natriuretic peptide (pg/mL) [a]		1.08	0.91–1.29	0.3833
Age (years)	813	1.02	1.01–1.04	0.0011
CRT		0.87	0.21–3.6	0.8416
CRT * Age (years)		1.00	0.97–1.02	0.6400
Ischaemic (yes/no)	812	1.68	1.29–2.19	0.0001
CRT		0.48	0.35–0.66	0.0001
CRT * Ischaemic (yes/no)		1.49	0.99–2.26	0.0583

Table 2.2 Potential predictors of risk: results of univariable analyses
a = $\log_e$ transformed, * denotes an interaction

Notes for Table 2.2
Mitral regurgitation represents the results of fitting single Cox Proportional Hazards model, a patient's time to the primary event being assumed to be dependent on mitral regurgitation and also the presence or absence of CRT. The term CRT * log(MR) is a treatment modifier, this means that the beneficial

effect of CRT may be reduced or increased depending on the patients level of mitral regurgitation. Mitral regurgitation is a significant predictor of outcome, $P < 0.0001$, however, the P-value for CRT * log(MR) > 0.05 so mitral regurgitation does not significantly change the benefit a patient may receive from CRT.

The most appropriate transformation of each variable is indicated (for example a logarithmic transformation led to the best model fit based on the AIC for MR). The remaining variables (beta-blocker use and QRS width) were not significantly associated with outcome and did not predict response to CRT. Those variables identified to be significantly ($P < 0.05$) associated with the primary composite outcome, time to death from any cause, or an unplanned hospitalization for a major cardiovascular event were entered in a multivariable Cox Proportional Hazards model (Table 2.3)

	Transformation	**Hazard ratio**	**95% CI**	**P-value**
Significant Predictors of overall outcome				
Mitral regurgitation	Log_e	1.71	1.38–2.12	0.0001
N-terminal pro-braina natriuretic peptide (pg/ml)	Log_e	1.31	1.17–1.47	0.0001
Systolic blood pressure (mmHg)	Linear	0.99	0.98–1.00	0.0698
Interventricular mechanical delay (ms)	Linear	1	0.99–1.01	0.7617
Aetiology (ischaemic) (yes/no)	Factor	1.89	1.45–2.46	0.0001
CRT (yes/no)	Factor	0.608	0.47–0.79	0.0003
Predictors of response to CRT				
Systolic blood pressure (mmHg)*CRT	Linear	1.02	1.00–1.03	0.0183
Interventricular mechanical delay (ms)*CRT	Linear	0.99	0.98–1.00	0.0084

Table 2.3 Significant Predictors of outcome and response to CRT

Notes for Table 2.3
Mitral regurgitation and N-terminal pro-brain natriuretic peptide have been identified as statistically significant predictors of outcome. The terms CRT * SBP and CRT * IVMD represent modifiers of response to CRT, i.e. both systolic blood pressure and interventricular mechanical delay may modify the beneficial effect of CRT. The P-values for CRT * SBP and CRT * IVMD are both < 0.05,indicating that systolic blood pressure and interventricular mechanical delay are statistically

significant. Note that individually systolic blood pressure nor interventricular mechanical delay are statistically significant, in other words they are not predictors of outcome. The P-value for CRT is relatively large (0.0347) due to the inclusion of the CRT modifiers in the model.

Ischaemic aetiology, more severe MR, and increased NT-pro-BNP were all independent predictors of time to death or unplanned or unplanned cardiovascular hospitalization irrespective of randomised treatment (CRT) (Hazard ratio (HR) 1.89, 95% CI 1.45 to 2.46, HR 1.71, 95% CI 1.38 to 2.12 and HR 1.31, 95% CI 1.17 to 1.47, respectively) and increasing SBP with a decreasing risk of an event (HR 0.99, 95% CI 0.98 to 1.00) (Figure 2.1A–E). Note, in Figures 2.1A-E refer to median values, for the combined data, i.e. the median for the treatment and control groups combined. The prognostic model for the CARE-HF data includes two interaction terms CRT*Interventricular Mechanical Delay and CRT*Systolic Blood Pressure. These interaction terms involve a continuous and a binary variable, orthogonalization of the continuous variables and re-coding of the binary variables can be of great help in interpreting interaction terms. A Continuous variable X is transformed in the following way $X - \overline{X}$, a binary variable $I(1,0)$ is re-coded as 0.5 and -0.5 . Table 2.2a presents the same univariate analyses as in Table 2.2 but continuous variables have been transformed as described above along with re-coding of binary variables. The hazard ratio for CRT is much more stable across the univariate models compared with those presented in Table 2.2

	n	Hazard ratio	95% CI	P-value
Mitral regurgitation [a]	605	1.807	1.511 - 2.610	<.0001
CRT		0.692	0.541 - 0.883	0.0031
CRT * Mitral regurgitation [a]		0.716	0.501 - 1.024	0.0670
Interventricular mechanical delay (ms)	735	0.989	0.985 - 0.992	<.0001
CRT		0.632	0.507 - 0.787	<.0001
CRT * Interventricular mechanical delay (ms)		0.992	0.985 - 1.00	0.0473
End-systolic volume index (mL/m2) [a]	732	1.515	1.138 - 2.018	0.0044
CRT		0.618	0.497 - 0.768	<.0001
CRT * End-systolic volume index (mL/m2) [a]		0.999	0.564-1.771	0.9978
Glomeruler filtration rate (ml/min/1.73 m2)	739	0.986	0.980 - 0.991	<.0001
CRT		0.611	0.489 - 0.764	<.0001
CRT * Glomeruler filtration rate (ml/min/1.73 m2)		0.997	0.986 - 1.008	0.5811
Systolic blood pressure (mmHg)	803	0.993	0.987 - 1.00	0.0364
CRT		0.631	0.513 - 0.775	<.0001
CRT * Systolic blood pressure (mmHg)		1.013	1.00 - 1.025	0.0491
Ejection fraction (%)[a]	745	0.422	0.270 - 0.659	0.0002
CRT		0.639	0.516 - 0.792	<.0001
CRT * Ejection fraction (%)[a]		1.242	0.509 - 3.028	0.6341
N-terminal pro-brain natriuretic peptide (pg/mL) [a]	732	1.534	1.403 - 1.676	<.0001
CRT		0.593	0.470 - 0.750	<.0001
CRT * N-terminal pro-brain natriuretic peptide (pg/mL) [a]		1.082	0.906 - 1.292	0.3833
Age (years)	813	1.021	1.011 - 1.032	<.0001
CRT		0.621	0.506 - 0.763	<.0001
CRT * Age (years)		0.995	0.974 - 1.016	0.6400
Ischaemic (yes/no)	812	2.058	1.671 – 2.534	<.0001
CRT		0.589	0.478 - 0.725	<.0001
CRT * Ischaemic (yes/no)		1.494	0.986 – 2.263	0.0583

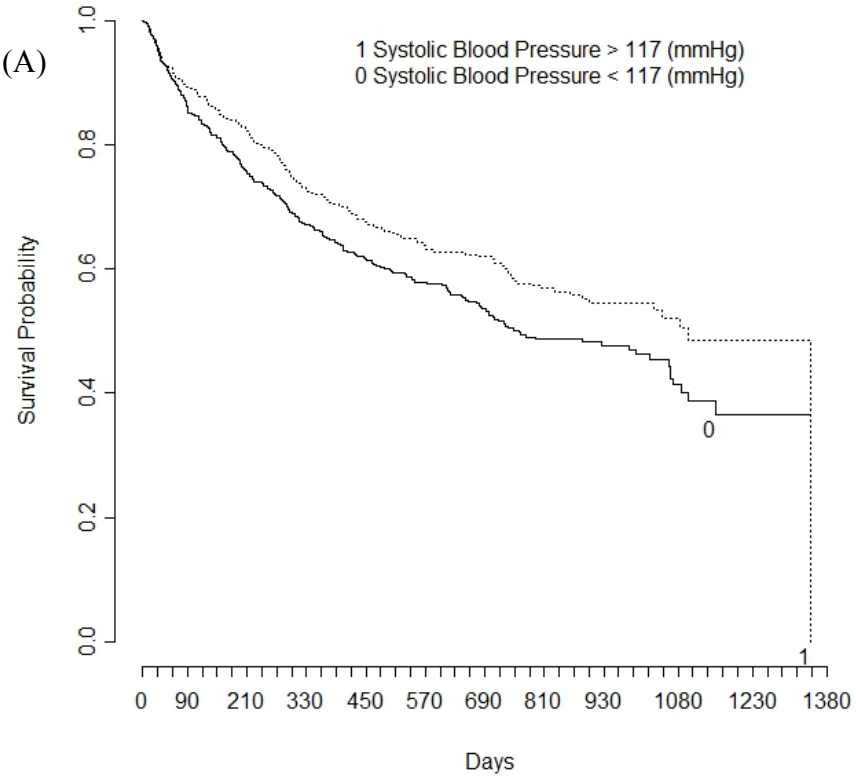
Table 2.2a Results of Orthogonalization Potential predictors of risk: results of univariable analyses
a = $\log_e$ transformed, * denotes an interaction

Notes for Table 2.2a

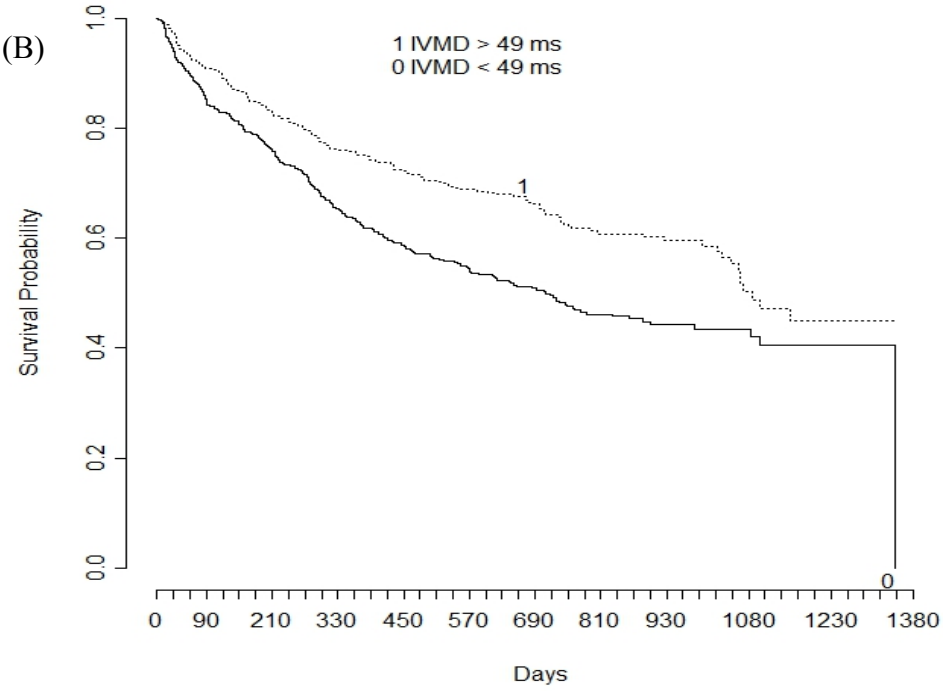
Mitral regurgitation represents the results of fitting single Cox Proportional Hazards model, a patient's time to the primary event being assumed to be dependent on mitral regurgitation and also the presence

or absence of CRT. The term CRT * log(MR) is a treatment modifier, this means that the beneficial effect of CRT may be reduced or increased depending on the patients level of mitral regurgitation. Mitral regurgitation is a significant predictor of outcome, $P < 0.0001$, however, the P-value for CRT * log(MR) > 0.05 so mitral regurgitation does not significantly change the benefit a patient may receive from CRT.

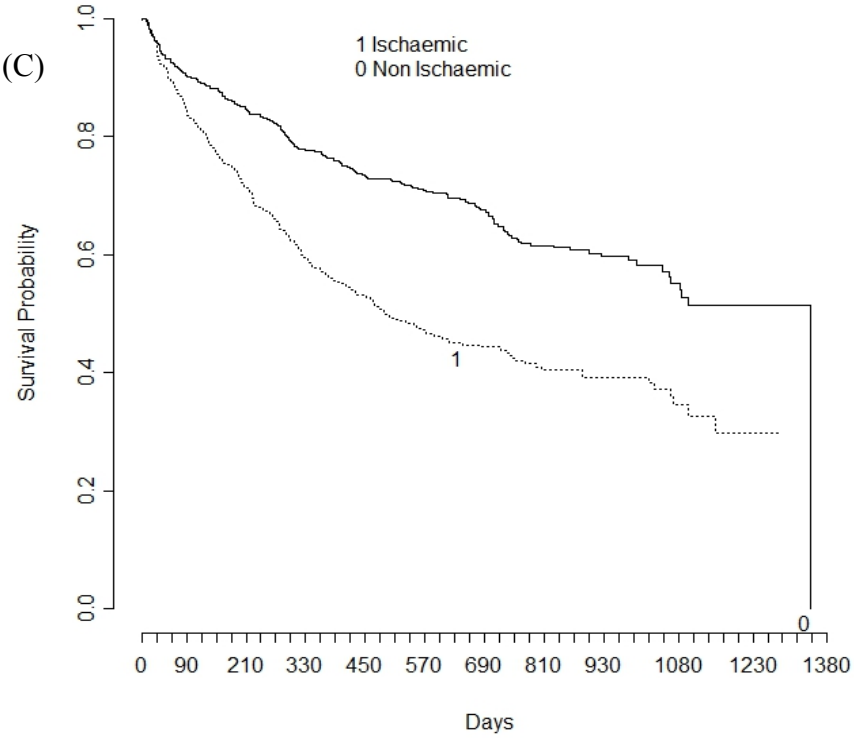
(A)



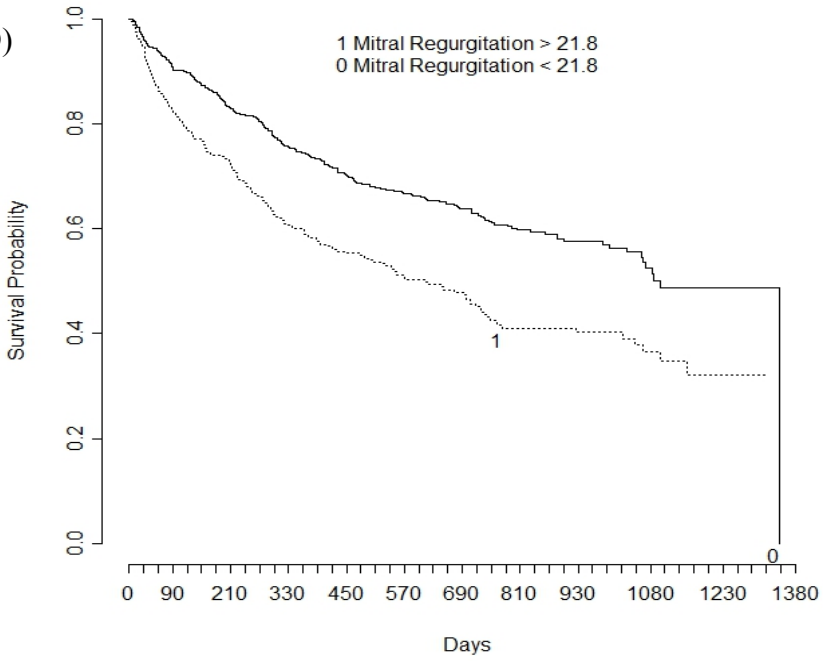
(B)


1 IVMD > 49 ms
0 IVMD < 49 ms
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
1
0
0 90 210 330 450 570 690 810 930 1080 1230 1380
Days

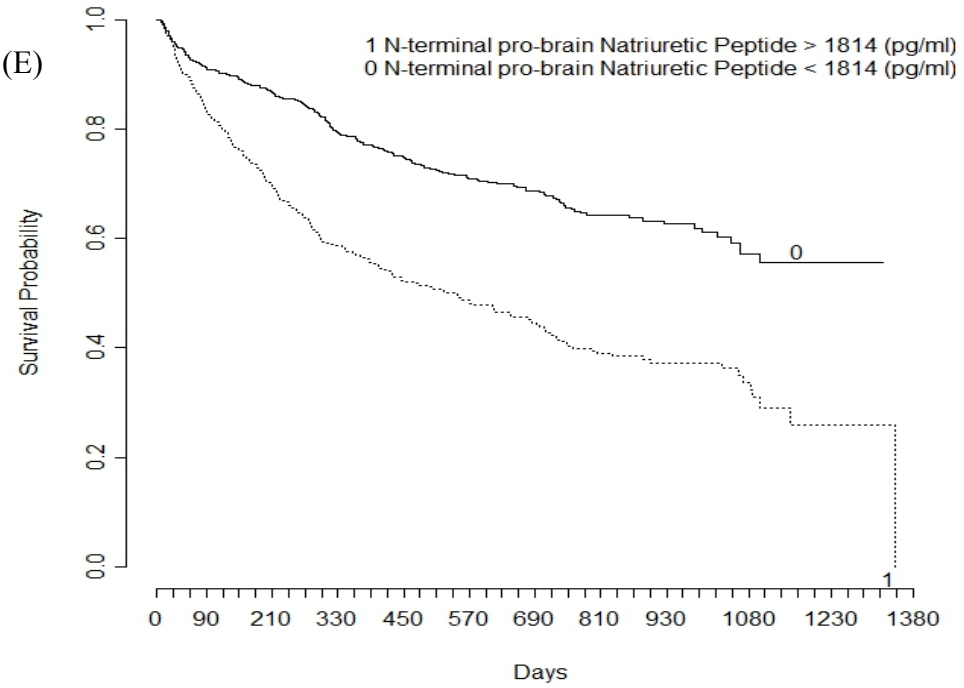
(C)


1 Ischaemic
0 Non Ischaemic
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
1
0
0 90 210 330 450 570 690 810 930 1080 1230 1380
Days

(D)


1 Mitral Regurgitation > 21.8
0 Mitral Regurgitation < 21.8
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
1
0
0 90 210 330 450 570 690 810 930 1080 1230 1380
Days

(E)


1 N-terminal pro-brain Natriuretic Peptide > 1814 (pg/ml)
0 N-terminal pro-brain Natriuretic Peptide < 1814 (pg/ml)
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1
0 90 210 330 450 570 690 810 930 1080 1230 1380
Days

(F)

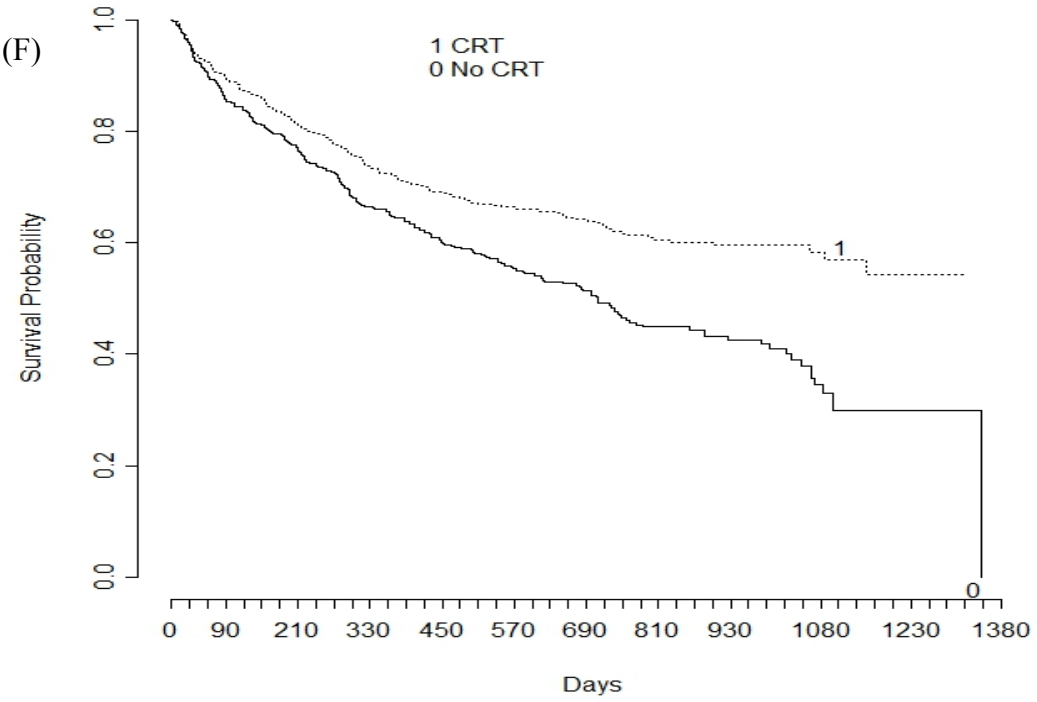


Figure 2.1 (A) Time to first primary event by systolic blood pressure. (B) Time to first primary event by interventricular mechanical delay. (C) Time to first primary event by aetiology (ischaemia). (D) Time to first primary event by mitral regurgitation. (E) Time to first primary event by N-terminal pro-brain natriuretic peptide (pg/ml).(F) Time to first primary event by Cardiac Resynchronisation

Only two variables, IVMD and SBP predicted response to CRT, with modest statistical precision (Figures 2.2 and 2.3). Patients with increasing SBP appear to receive reduced benefit from CRT (HR 1.02, 95% CI 1.00–1.03), whereas those patients with more severe IVMD appear to benefit more from treatment (HR 0.99, 95% CI 0.98–1.00).

It is important that the validity of the proportional hazards assumption is assessed. The following SAS code shows how the ASSESS option in PHREG is used to test the proportional hazards assumption for the final model.

```
ods graphics on;
   proc phreg data=card.progex3;
      class Ischemic treat /desc;
      model futime*primary(0)= treat mitral_r Roche supsys IVMD
Ischemic trsup trivm;
      assess PH/  resample seed=7548;
      run;
ods graphics off;
```

In the above code assess PH specifies that proportional hazards assumption are tested, Table 2.4 shows the Kolmogorov type supremum test for proportional hazards produced by ASSESS, ASSESS uses the methods of Lin (Lin, Wei & Ying 1993). From Table 2.4 there is some evidence that the proportional hazards assumption is violated for CRT (p=0.0380). The remaining variables appear not to violate the proportional hazards assumption. The non proportional hazards for CRT could be dealt with by fitting a model with a time dependent variable, this could be achieved by introducing the term CRT*$\log_e$(time) .The time dependent variable must be defined after the model statement in PHREG .The results of fitting this model are shown in Table 2.5. Since the main objective of the model presented in this thesis is to identify modifiers of CRT and not to determine the effect of CRT itself, it might be argued that the non proportional hazards for CRT could be ignored and that the model presented in Table 2.3 would be adequate for the purposes of identifying modifiers of CRT. Another approach to accommodating non proportional hazards would be to develop a stratified model, the strata being the variable for which proportional hazards is violated. This approach is valid if the stratification is based on a variable which is not of primary interest.

Supremum Test for Proportional Hazards Assumption

Variable	Maximum Absolute Value	Replications	Seed	p
CRT	1.4477	1000	7548	0.0380
Mitral regurgitation	0.9270	1000	7548	0.3620
N-terminal pro-brain natriuretic peptide (pg/ml)	0.7351	1000	7548	0.6200
Systolic blood pressure (mmHg)	0.9699	1000	7548	0.2390
Interventricular mechanical delay (ms)	0.8972	1000	7548	0.5110
Aetiology (ischaemic Y/N)	0.9108	1000	7548	0.3930
Systolic blood pressure (mmHg)*CRT	0.4964	1000	7548	0.9030
Interventricular mechanical delay (ms)*CRT	1.0169	1000	7548	0.3300

Table 2.4 Test of Proportional Hazards (Note Mitral regurgitation and N-terminal pro-brain natriuretic peptide (pg/ml) are $\log_e$ transformed)

	Parameter Estimate	Standard Error	Chi-Square	p	Hazard Ratio
CRT	0.69621	0.58332	1.4245	0.2327	2.006
Mitral regurgitation	0.54294	0.10870	24.9468	<.0001	1.721
N-terminal pro-brain natriuretic peptide (pg/ml)	0.27144	0.05912	21.0796	<.0001	1.312
Systolic blood pressure (mmHg)	-0.0002648	0.00369	0.0051	0.9428	1.000
Interventricular mechanical delay (ms)	-0.00528	0.00255	4.2932	0.0383	0.995
Aetiology (ischaemic Y/N)i	0.62633	0.13515	21.4765	<.0001	1.871
Systolic blood pressure (mmHg)*CRT	0.01723	0.00727	5.6161	0.0178	1.017
Interventricular mechanical delay (ms)*CRT	-0.01202	0.00497	5.8433	0.0156	0.988
CRT*$\log_e$ (time)	-0.22803	0.10857	4.4116	0.0357	0.796

Table 2.5 Model with time dependent variable CRT*$\log_e$ (time) (Note Mitral regurgitation and N-terminal pro-brain natriuretic peptide (pg/ml) are $\log_e$ transformed)

.

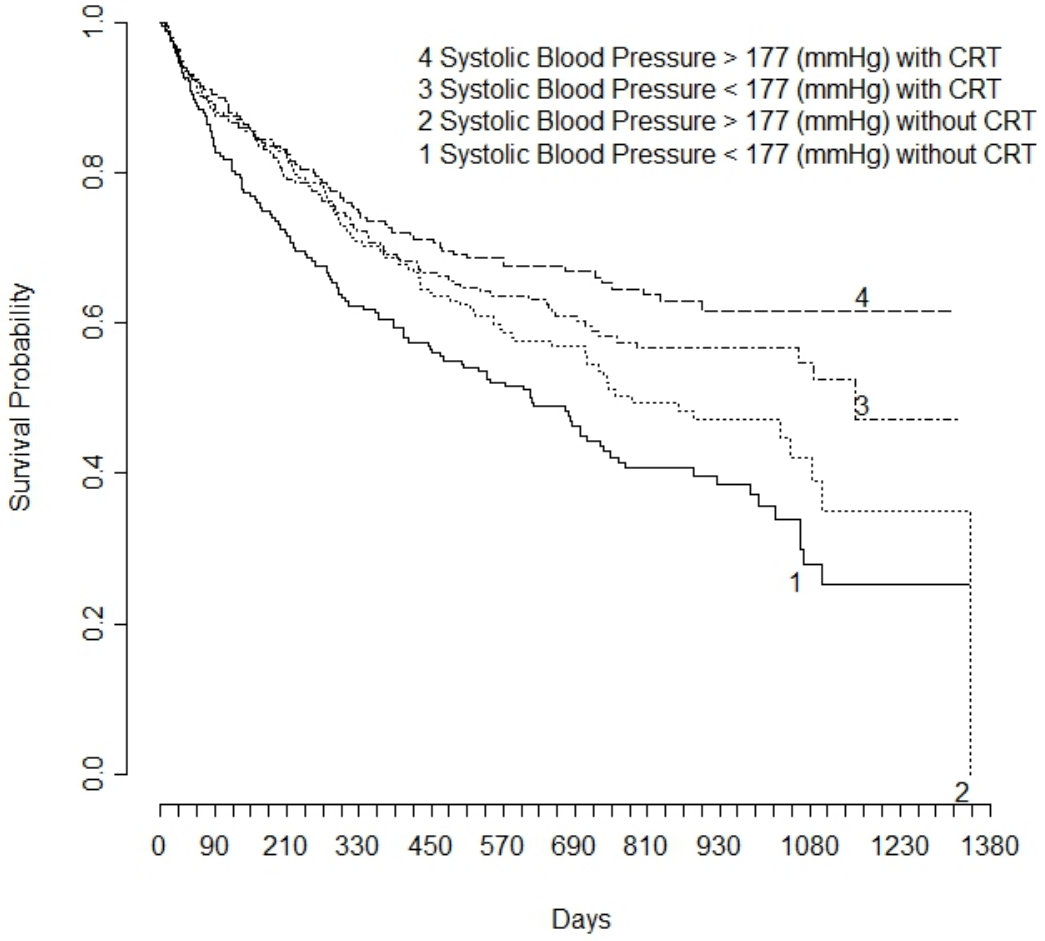


Figure 2.2 Time to first primary event by systolic blood pressure (mmHg) and cardiac resynchronization therapy.

Number at risk Systolic Blood Pressure (SBP)				
SBP	1 Month	3 Months	6 Months	12 Months
<117 (mmHg) without CRT	198	173	154	125
<117 (mmHg) with CRT	193	179	166	141
>117 (mmHg) without CRT	186	171	161	133
>117 (mmHg) with CRT	191	181	167	147

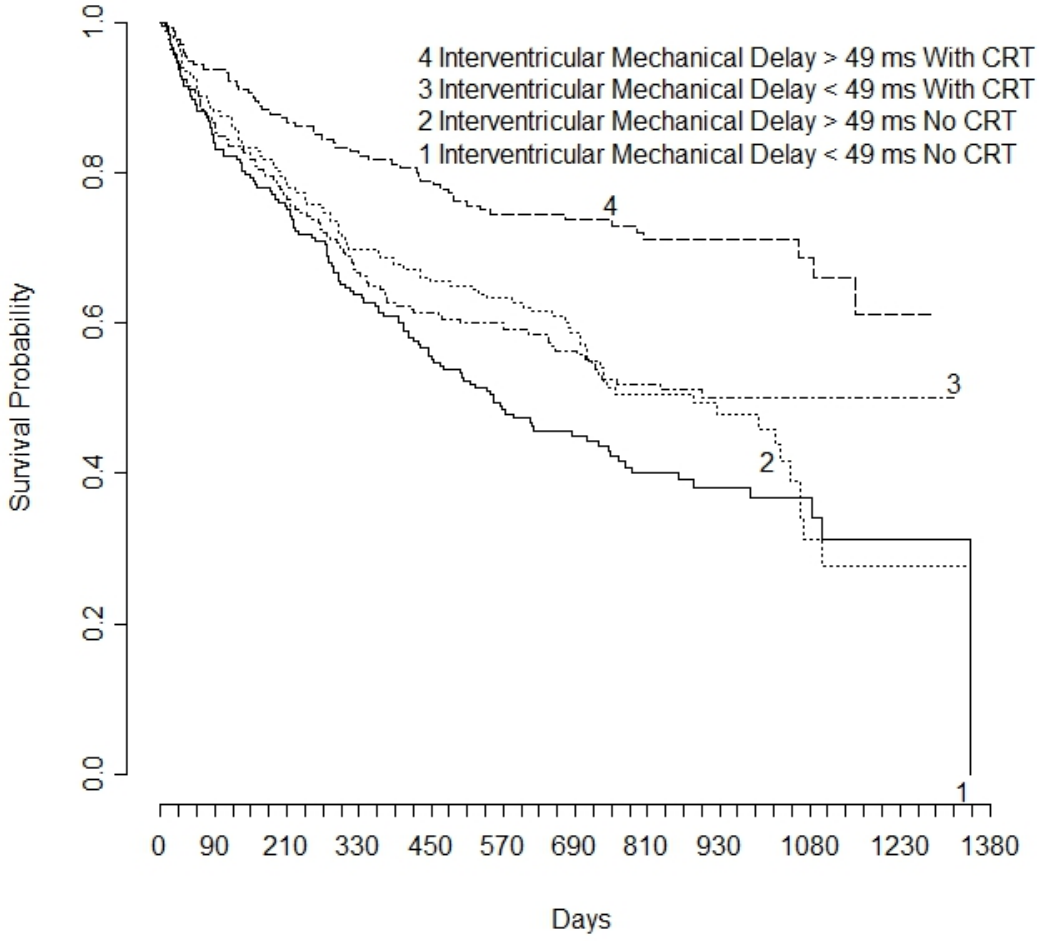


Figure 2.3 Time to first primary event by interventricular mechanical delay (ms) and cardiac resynchronization therapy.

Number at risk Interventricular Mechanical Delay (IVMD)				
IVMD	1 Month	3 Months	6 Months	12 Months
<49 ms without CRT	202	179	164	130
<49 ms with CRT	213	194	179	145
>49 ms without CRT	181	165	151	128
>49 ms with CRT	176	169	159	147

2.2.1 Discussion

The CARE-HF trial demonstrated that CRT exerts a substantial reduction in morbidity and mortality with little evidence of heterogeneity in pre-defined subgroups (Cleland *et al.* 2005). This more detailed analysis provides evidence that IVMD and to a lesser extent SBP predict a patients' response to CRT. These finding must be treated with a degree of caution as the model is exploratory and the interactions between CRT and either IVMD or SBP were not strong. However, the observed interaction between IVMD and the effects of CRT are consistent with the view that IVMD is a more precise physiological marker of cardiac dyssynchrony, the problem that CRT is designed to treat, than any other variable analysed. IVMD could therefore potentially be used as an inclusion criterion in future randomized controlled trials examining the effects of CRT in patient populations not included in CARE-HF, such as patients with less severe symptoms or with shorter QRS intervals. Whether IVMD should now be used in preference or in addition to QRS duration to identify whether a patient should receive CRT is a matter for the individual clinician to decide and for future research. It is of great importance to note that IVMD is the best predictor of response to CRT in a population having large volumes, low EF, and broad QRS. We cannot state that IVMD is a better predictor of response to CRT in other populations (Ghio *et al.* 2004).

Patients recruited to the study had severe heart failure (NYHA class III–IV) and therefore had an inherently high risk of experiencing the primary outcome during the study follow-up (which ranged from 18 to 44.7 months). The hazard functions from the model are based upon prediction of event rates across the maximum follow-up

from the study, which had reached 55% in the control group a mean 29.4 months of follow-up. In order to estimate the absolute risk of an event with changing SBP and IVMD, the remaining clinical predictors were held constant. It is important to note that since these are also strong clinical predictors of outcome changing these values from the median has a large impact on the estimates of absolute risk. For example, in a non-ischaemic patient not receiving CRT with a SBP of 117 mmHg, use of lower interquartile range values for mitral regurgitation and NT-pro-BNP results in an estimate of absolute risk of approximately 0.84, an absolute reduction of around 13%. The plasma concentration of NT-pro-BNP was a strong predictor of clinical outcome. Other competing measures of ventricular dysfunction were eliminated from the multivariable model. CRT reduces the severity of mitral regurgitation and plasma concentrations of NT-pro-BNP, and CRT has substantial clinical benefits in a broad range of patients with evidence of cardiac dyssynchrony, poor LV systolic function, and persistent symptoms despite pharmacological therapy. This analysis provides further evidence that a measure of cardiac dyssynchrony rather than the QRS interval on the ECG is currently the best marker of dyssynchrony. However, the predicted benefits from the model indicate that CRT appears worthwhile across the range of patients included in the CARE-HF trial. In the next chapter we will consider the function form of the model, and how the correct form can be specified by use of cubic splines or fractional polynomials.

CHAPTER 3 RISK ESTIMATION

- Relative and absolute risk
- Risk score produced using the prognostic model for the CARE-HF data
- Risk score calculators for the CARE-HF data are presented

3.0.0 Introduction

The purpose of a prognostic model is to aid clinical decision making (Wyatt & Altman 1995). A prognostic model can enable a doctor to assess risk for an individual patient. Prognostic models can be used by a doctor to assist in making an informed and rational choice as to what treatment a patient should or should not receive. For example it may be that several treatments are available, by using a prognostic model a doctor can determine the treatment that will offer maximum benefit to the patient. Some treatments may be very costly and unfortunately due to financial constraints it may be necessary to target resources at those patients who are most likely to respond positively to a particular treatment regime. Bodies such as NICE (NICE 2009) might rely on a prognostic model in targeting resources. A doctor often needs to determine the risk (Sedgwick 2001) that a patient will experience some event of interest. For instance given a patient' age, weight, blood pressure and the fact that the patient is receiving treatment for a heart condition, what is the chance of the patient suffering a heart attack? When considering risk for an individual patient the term absolute risk is employed. If instead of individual patients groups of patients are considered e.g.

patients receiving treatment versus those not receiving treatment, male versus female patients; then the term relative risk is employed. Relative risk is a comparison of the risk of some event of interest occurring in two groups of patients. The fact that a prognostic model will be used in the 'real' world to guide a clinician in making important decisions emphasises the need for good quality models. Also the model needs to be available in a form that is easily used by a clinician to calculate risk. The prognostic model can be made available to the clinician as a risk score calculator. A risk score calculator is an implementation of the prognostic model in software form. The EuroScore (Euroscore Website) calculator is an example of a risk score calculator. I have produced two simple risk score calculators using the prognostic model for the CARE-HF data (Richardson *et al.* 2007). The calculators allow the clinician to quickly and easily calculate a risk score for an individual patient. The risk score gives a measure of how likely a patient is to die from any cause or be hospitalised due to a major cardiovascular event. The calculators could be used on a computer system at a GP's surgery, or if installed on a laptop computer or hand held device could be used in a bedside prognosis in a hospital ward or a patient's home. Figure 3.1 below shows the risk score calculator produced by the present author running on Microsoft Windows XP. The calculator was written using Visual Basic For Applications (VBA) and is embedded in a Microsoft Excel workbook. Figure 3.2 shows the risk score calculator running on GNU/Linux, this version of the calculator was written using Gambas (Benoît Minisini Website 2009) a free software equivalent to Visual Basic and is a standalone program.

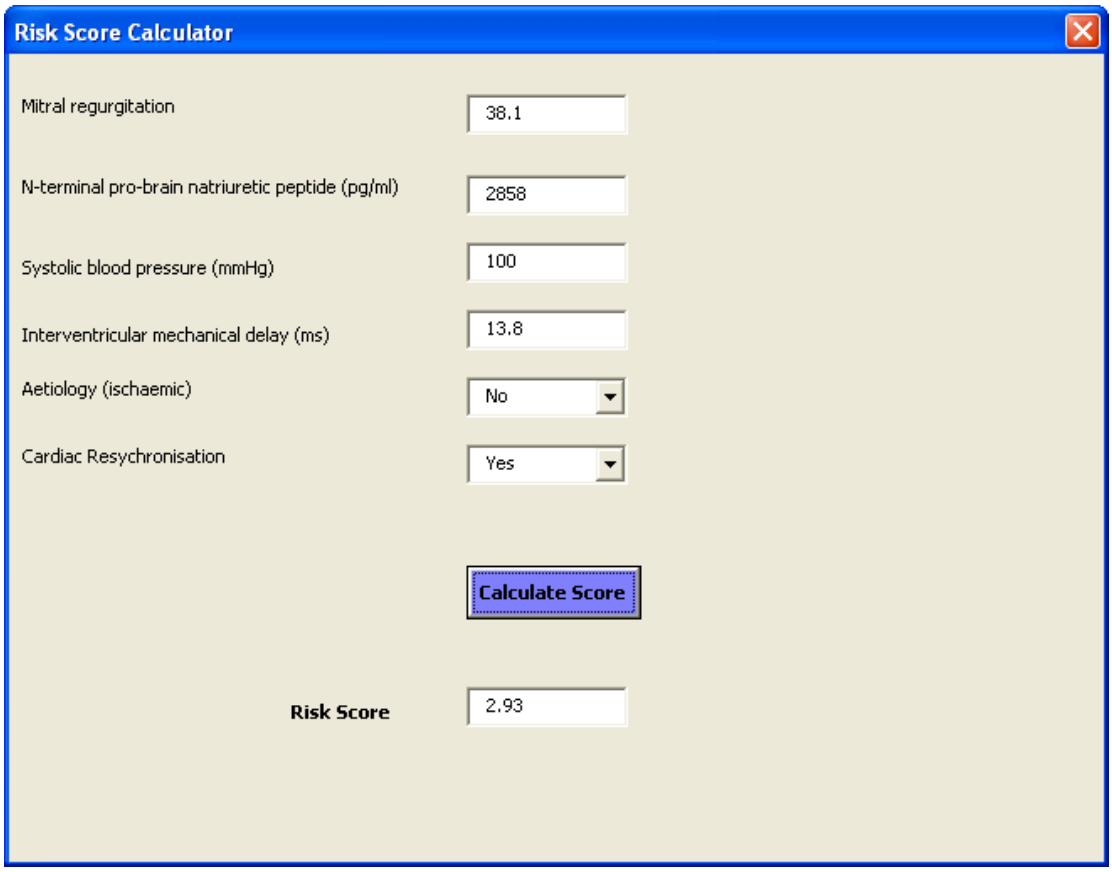


Figure 3.1 Risk Score Calculator developed by the author running on Microsoft Windows XP

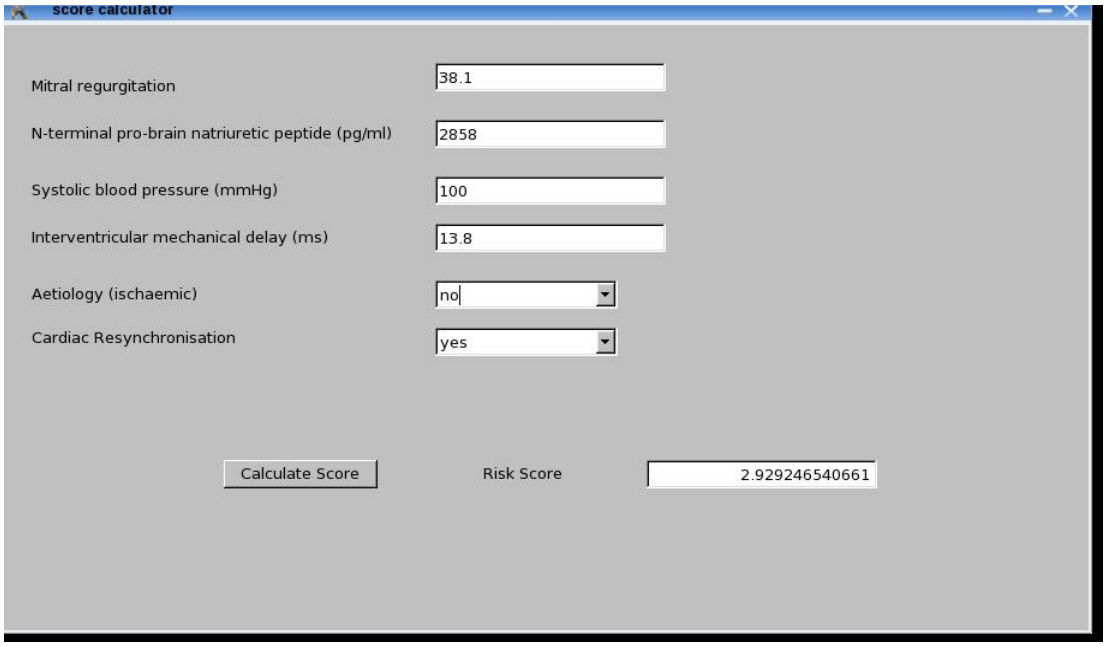


Figure 3.2 Risk Score Calculator developed by the author running on GNU/Linux

3.1.0 Calculation of risk scores

Once a prognostic model has been developed it is possible to determine both absolute and relative risks. Also the prognostic model can be used to generate a risk score, this risk score is the linear predictor $\eta = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$. For the CARE-HF model the risk score does not include β_0.To illustrate how these estimates of risk are obtained I shall use the prognostic model developed for the CARE-HF data (Richardson *et al.* 2007).

The coefficients of the final model can be used to generate a risk score for an individual patient. A quick and convenient way of estimating risk for an individual patient is to substitute patient characteristics in the Cox Proportional Hazards model. An example showing how the risks score is calculated as follows:

Risk score for patient with mitral regurgitation of 38.1, NT-pro-BNP of 2858 pg/ml, systolic blood pressure of 100 mmHg, IVMD of 13.8 ms, ischaemic, and in receipt of CRT would be calculated as follows:

Risk Score

$$= 0.5379\log_e(MR) + 0.2717\log_e(NT - pro - BNP) - 0.0001SBP - 0.0055IVMD$$

$$+ 0.6340ischaemic + 0.0172(CRT * SBP) - 0.0131(CRT * IVMD) - 0.4978CRT$$

So for the patient above we would have Risk Score

$$= 0.5379(\log_e(38.1) - 2.94) + 0.2717(\log_e(2858) - 7.43) - (0.0001 \times (100 - 117)) + (0.0055 \times (13.8 - 49.9)$$

$$+ (0.6340 \times 0.5) + 0.0172(0.5 \times (100 - 117)) - 0.0131(0.5 \times (13.8 - 49.9)) - (0.4978 \times 0.5) = 0.48$$

Figure 3.3 shows a plot of the risk score versus the probability of experiencing the primary event. By using the predict option in PHREG the survivor function estimate s

can be obtained, a new data set containing the risk score and the probability of experiencing the primary event (1-s) before the end of the follow up period can then be created.

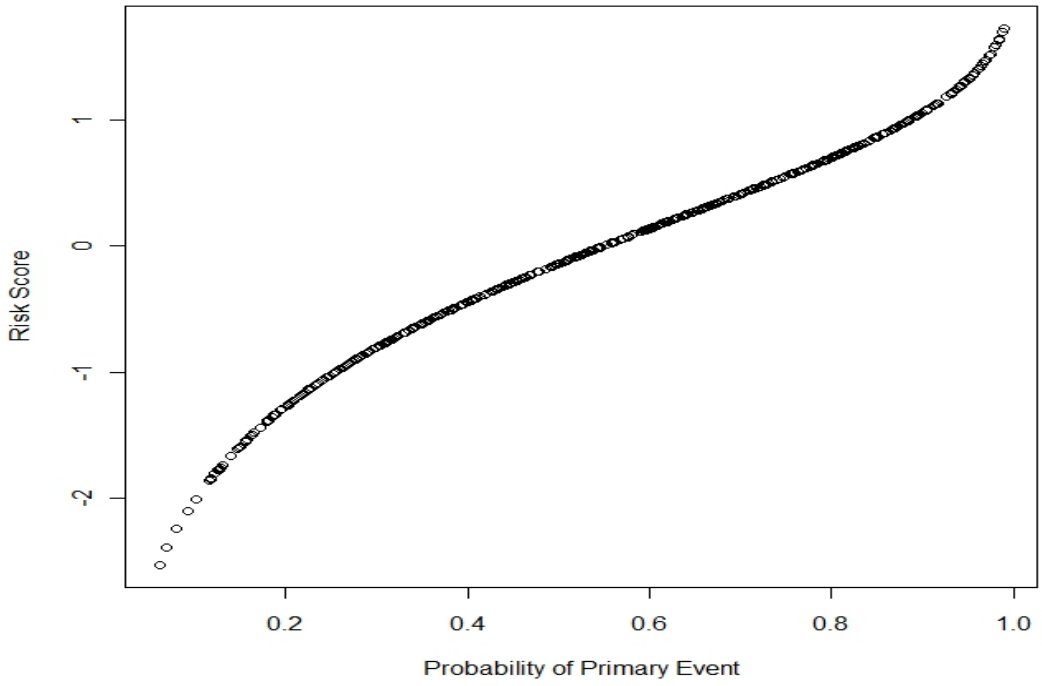


Figure 3.3 Risk score vs. probability of primary event before end of follow-up period.

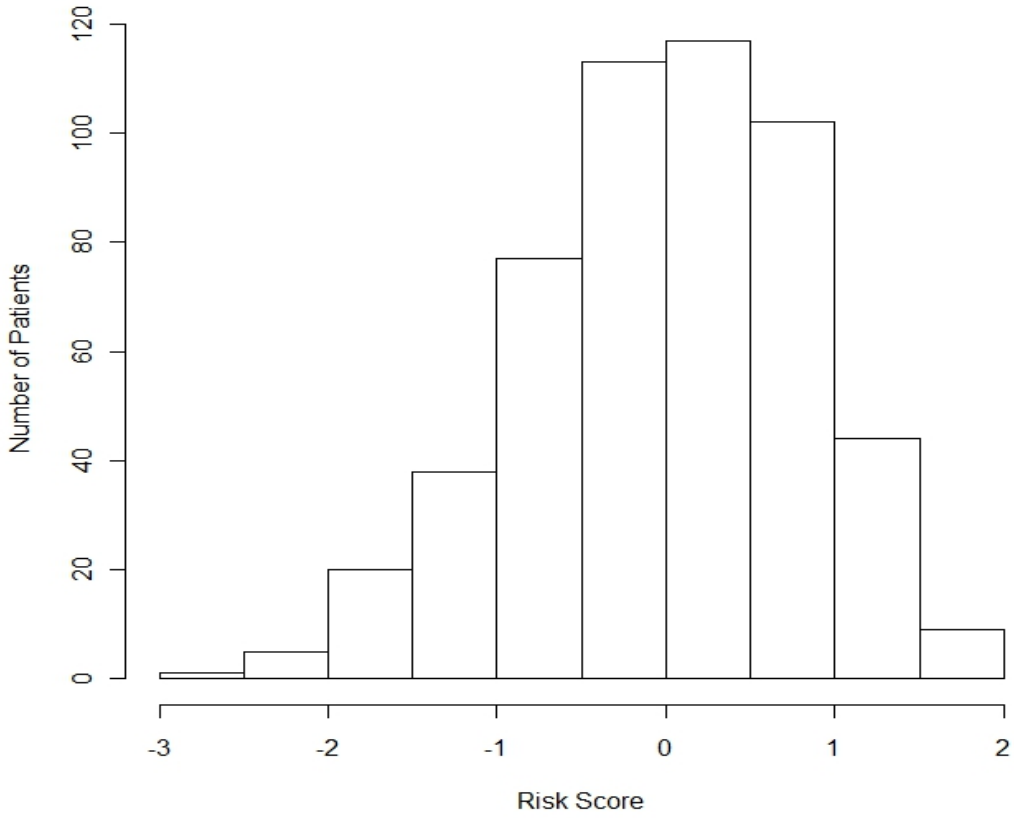


Figure 3.4 Histogram of risk score for patients before end of follow-up period.

3.2.0 Estimation of absolute risk

Estimates of the survival function $S(t)$ and the absolute risk $1 - S(t)$ were produced using the SAS procedure PHREG . Estimation of absolute risk using real patient data provides clinically relevant estimates of risk. Risk estimates were derived on the basis of the maximum follow-up in the CARE-HF study, which was 44.7 months, although including censorship patients were only followed for on average 29.4 months. Thus predicted event rates are considerably higher than those actually observed in the trial. The effect of SBP and IVMD on the absolute risk of a patient experiencing death from any cause or an unplanned

hospitalization for a major cardiovascular event in the presence and absence of CRT or ischaemic heart disease are shown in Tables 3.1 and 3.2, respectively. In both examples, mitral regurgitation, NT-pro-BNP, and IVMD were held constant at the median values (see Table 3.3) the values NT-pro-BNP and IVMD given in Tables 3.1 and 3.2 are those after othogonalization . The estimated absolute risk of experiencing death or an unplanned hospitalization for cardiovascular cause for a non-ischaemic patient with a SBP of 117 mmHg (the median for the whole dataset) on medical therapy (but not CRT) was 0.62 over the entire trial duration (Table 3.1). Treatment of such a patient with CRT reduces the estimated absolute risk to 0.44. The presence of ischaemia led to an increase in absolute risk to 0.67 and 0.84 in the presence and absence of CRT, respectively. The absolute risk of experiencing event decreased with increasing SBP, this is due to the fact that although increased SBP alone is associated with a decrease in risk, the statistical interaction between SBP and CRT is associated with a small increase in risk. The absolute risk for a patient with IVMD of 49 ms vs. a patient with IVMD of 66 ms in the presence and absence of ischaemia and CRT is shown in Table 3.2. Increasing the IVMD from 49 to 66 ms leads to an increase in the absolute risk of experiencing an event, this result contradicts what would be expected from the model given that the coefficient for IVMD is –ve. From figure 2.3 it appears that increasing IVMD does diminish risk, however at around 1050 days the survival curves for patients not in receipt of CRT start to cross. The patients considered in Tables 3.1 and 3.2 had survived beyond 1050 days. . The estimated absolute risk of experiencing a primary outcome event may seem surprisingly high in some cases (absolute risk of 0.99 as shown in Tables 3.1 and 3.2). However, patients recruited to the study had severe heart failure (NYHA class III–IV) and therefore had an inherently high risk of experiencing the primary outcome during the study follow-up

(which ranged from 18 to 44.7 months). The hazard functions from the model are based upon prediction of event rates across the maximum follow-up from the study, which had reached 55% in the control group in mean 29.4 months of follow-up. In order to estimate the absolute risk of an event with changing SBP and IVMD, the remaining clinical predictors were held constant. It is important to note that since these are also strong clinical predictors of outcome changing these values from the median has a large impact on the estimates of absolute risk.

Patient	Systolic Blood Pressure (mmHg)	Aetiology (Ischaemic)	Cardiac Resynchronisation Therapy	Absolute Risk
1	-0.49	No	Yes	0.44
2	-0.49	No	No	0.62
3	-0.49	Yes	Yes	0.67
4	-0.49	Yes	No	0.84
5	12.5	No	Yes	0.48
6	12.5	No	No	0.58
7	12.5	Yes	Yes	0.71
8	12.5	Yes	No	0.81

Table 3.1 Estimated absolute risk of an event for patients with different systolic blood pressures (117–130 mmHg) with and without cardiac resynchronisation therapy and in the presence and absence of ischaemic heart disease.

Patient	Interventricular Mechanical Delay (ms)	Aetiology (Ischaemic)	Cardiac Resynchronisation Therapy	Absolute Risk
1	0.1	No	Yes	0.44
2	0.1	No	No	0.62
3	0.1	Yes	Yes	0.67
4	0.1	Yes	No	0.84
5	17.44	No	Yes	0.38
6	17.44	No	No	0.63
7	17.44	Yes	Yes	0.59
8	17.44	Yes	No	0.85

Table 3.2 Estimated absolute risk of an event for patients with varying interventricular mechanical delay (49–66 ms) with and without cardiac resynchronisation therapy and in the presence and absence of ischaemia .

	Control			**Treatment**		
	n	**median**	**(IQR)**	**n**	**median**	**(IQR)**
Age (years)	403	66	(59–72)	409	67	(60–73)
Aetiology (ischaemic Y/N)	Y=153 N=250			Y=186 N=223		
Systolic blood pressure (mmHg)	399	110	(100–125)	404	110	(100–125)
Glomerular filtration rate (mL/min/1.73m^2)	372	61	(46–73)	367	60	(46–73)
N-terminal pro-brain natriuretic peptide (pg/ml)	370	1806	(719–3949)	362	1920	(744-4288)
Use of beta-blockers (Y/N)	Y=288 N=116			Y=298 N=111		
QRS width (ms)	394	160	(152–180)	401	160	(152–180)
Interventricular mechanical delay (ms)	370	50	(30–66)	365	49	(32–67)
End-systolic volume index (mL/m^2)	376	117	(94–147)	356	121	(92–151)
Ejection fraction (≤ 35%)	378	25	(22–29)	367	25	(21–29)
Mitral regurgitation	303	23	(11–34)	302	21	(12–33)

Table 3.3 Baseline characteristics of the patients,total number in study 813, IQR, interquartile range. [a] Mitral regurgitation defined as area of colour flow Doppler regurgitant jet divided by area of left atrium in systole, both in square centimetre.

3.3.0 Obtaining Estimates of Absolute Risk

It is worth commenting on what is involved in producing estimates of absolute risk using PHREG. The following steps are needed

1. Create a dataset containing a subset of example patients
2. Run PHREG with the baseline option
3. Create a dataset containing the absolute risk estimates

Step 1 can be accomplished using for example the following SAS code

```
data card.mrisks;
     input lmit lroc supsys IVMD ischemic trsup trivm treat;
     datalines;
      0.14 0.70 -0.49 0.100 0.5 -0.245 0.05 0.5
      0.14 0.70 12.5 0.100 0.5 6.25 0.05 0.5
      0.14 0.70 -0.49 0.100 -0.5 -0.24 0.05 0.5
      0.14 0.70 12.5 0.100 -0.5 6.25 0.05 0.5
      0.14 0.70 -0.49 0.100 0.5 0.245 -0.05 -0.5
      0.14 0.70 12.5 0.100 0.5 -6.25 -0.05 -0.5
      0.14 0.70 -0.49 0.100 -0.5 0.245 -0.05 -0.5
      0.14 0.70 12.5 0.100 -0.5 -6.25 -0.05 -0.5
;
```

Here I specify the names of the input variables and then construct a data set containing the example patients.

These patients have, lmit= $\log_e$(Mitral regurgitation) lroc= $\log_e$(N-terminal pro-brain natriuretic peptide), and Interventricular mechanical delay all held constant (set to the median) . Systolic blood pressure blood pressure is allowed to vary, as is aetiology (ischaemic), I compare the treatment and control groups. Step 2 is illustrated with the following code snippet

```
proc phreg data=card.valmod;
       model futime*primary(0)=lmit lroc supsys IVMD ischemic trsup
trivm treat/RL;
       baseline covariates=card.mrisks out=card.PredFin
survival=S/nomean;
run;
```

PHREG performs analysis of time to event data based on the Cox proportional hazards model. The survival time for each patient is assumed to follow its own hazard function $h_i(y)$, $h_i(y) = h_0(y)\exp(\tilde{X}_i\tilde{\beta})$, where $h_0(y)$ is an arbitrary and unspecified baseline hazard function. The survivor function $S(y, \tilde{X}_i)$ can be written as

$S_0(y)^{\exp(\tilde{X}_i\tilde{\beta})}$, where $S_0(y) = \exp(-\int_0^y h_0(u)du)$.

The BASELINE option in PHREG results in a new SAS data set that contains baseline function estimates for the variables listed in the SAS data set card.mrisks. In the above SAS code the survivor function $S(t)$ is estimated by the Breslow estimator (Breslow 1972) which is based on the empirical cumulative hazard function, alternatively the product limit estimator can be used (Kalbfleisch & Prentice 1980).

I can specify an out put dataset which will contain these estimates, (out=card.Predfin). The survival=S option means that I will obtain an estimate of the survivor function $S(t)$. Finally a dataset containing the estimates of absolute risk can be generated using the following SAS code

```
data card.absrisk;
     set card.PredFin;
       rsk=1-s;
run;
quit;
```

Here the estimate of $S(t)$ contained in the dataset PredFin is used to generate the estimate of absolute risk (rsk=1-s) which is contained in the dataset absrisk.

3.4.0 Which Measure of Risk Should Be Used?

A patient waiting in hospital for an operation would naturally want to know what is the benefit of undergoing surgery, he or she would want to know by how much would their risk (in the extreme case) of dying, be reduced . When considering a measure of risk reduction is there a benefit to using one measure as opposed to another? Chao (Chao *et al.*. 2003) discusses the issue of whether reporting relative risk reduction, absolute risk reduction, absolute survival benefit, or number needed to treat had an effect on a individuals decision to recommend that their mother undergo

chemotherapy (a hypothetical situation). Chao *et al.* found that the way in which risk reduction was presented does have a bearing on such a decision. They found that when an individual was presented with a relative risk reduction they were more likely to choose chemotherapy. Which measure of risk reduction to present is seemingly dependent on the patient's understanding of terms such absolute and relative risk. How one best presents risk to a patient is a very difficult question, I honestly do not believe that I can supply a definite answer to this question

CHAPTER 4 CUBIC SPLINES AND FRACTIONAL POLYNOMIALS

- Application of transformations such as the natural logarithm or the square root to the independent variables may lead to improved model fit
- More complex relationships can be modelled using cubic splines or fractional polynomials
- Restricted cubic spline applied to CARE-HF data

4.0.0 Introduction

In this chapter I will look at the use of cubic splines and fractional polynomials in developing a statistical model. The use of cubic splines and fractional polynomials is motivated by consideration of the adequacy of the functional form of the relationship between the dependent variable Y and the independent variable X . Unfortunately the real world confronts us with situations where Y is not related to X in a simple manner. A good example of this is the relationship between body mass index and risk of mortality for cardiac surgery see (Pagano *et al.* 2009). Pagano *et al* use cubic splines to model the relationship between body mass index and risk of mortality for cardiac surgery, see figure 4.1.

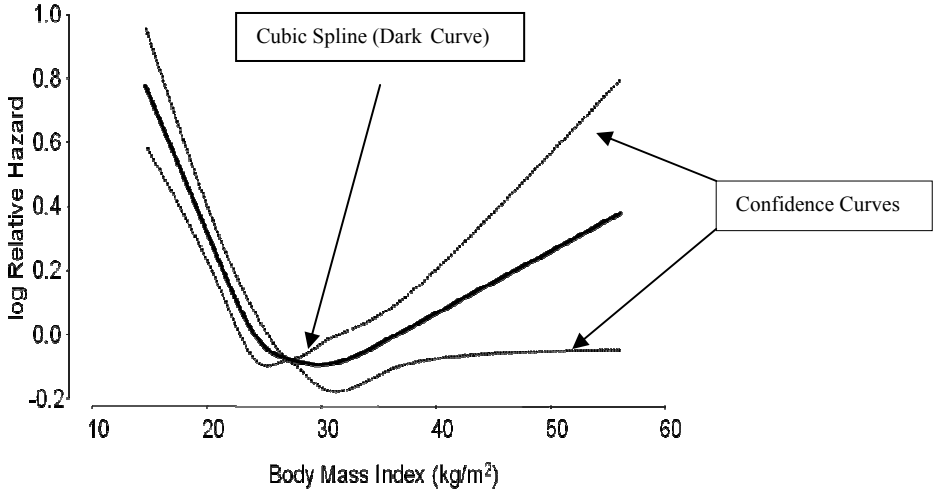


Figure 4.1 Non-linear (cubic spline) relationship between body mass index and risk of mortality for cardiac surgery. Adapted from (Pagano *et al.* 2009).

It might be assumed for instance that a simple linear relationship $y = mx + c$ is appropriate, but the data then leads the researcher to formulate a more complex model. One factor that will determine how well a model fits the data is the functional form of the relationship between Y and X. In developing a model, the researcher may make use of transformations of the independent variables in order to improve the fit, a typical example of such a transformation would be to consider $y = m\log_e x + c$. Amongst the other standard transformations are $\sqrt{x}, \frac{1}{x}$. It can be argued that it is natural to assume a linear relationship; if this proves not to be adequate then one might then consider taking the natural logarithm or the square root. Once the simple transformations have been applied then use of the cubic spline or fractional polynomial should be considered. The 'best' functional form may be quite complex and not easily obtained through analytic means, in this case numerical methods are used to approximate the relationship between Y and X. One such method is the cubic spline. I shall now look at some of the basic theory relating to cubic splines.

Before computer aided drawing software was available Engineers and Draughtsmen relied on a thin flexible rod called a spline. The spline was used to construct a curve through a series of points. The spline was anchored to the drawing board, and a number of weights were attached to the spline. The weights could then be moved and so the spline could be adjusted to obtain the best fit curve through the specified points. In mathematical terms, a spline is an approximation of a curve. A spline is an example of polynomial interpolation, or more correctly piecewise polynomial interpolation. Interpolation is the process of approximating some function $f(x)$ for x, where x is in

the interval (x_0, x_n). In polynomial interpolation we aim to find a polynomial $p(x)$ of degree n or less, such that $p(x_0) = f(x_0), p(x_1) = f(x_1),..., p(x_n) = f(x_n)$. In the literature $p_n(x)$ is used instead of $p(x)$ and f_n instead of $f(x_n)$.The polynomial $p_n(x)$ is known as an interpolation polynomial. There are various approaches to polynomial interpolation, for example Lagrange interpolation, Newton's Divided Difference interpolation, and Spline interpolation. It may be helpful to look at the Lagrange method (Box 1) in order to appreciate the general principles of interpolation and also to identify possible problems. In my discussion of Lagrange interpolation and splines I follow the derivations and notation found in Kreyszig (Kreyszig 1993), note an excellent explanation of splines can be found in Kreyszig's book.

Box 1 Lagrange Interpolation

Lagrange interpolation uses the following approach.

Assuming we have $(x_0, f_0),(x_1, f_1),...(x_n, f_n)$ (the point (x_i, f_i) is known as a node) then we can approximate the analytic function $f(x)$ by $p_1 = L_0(x) f_0 + L_1(x) f_1$, where $L_0(x) = \frac{x - x_1}{x_0 - x_1}$ and $L_1(x) = \frac{x - x_0}{x_1 - x_0}$. Notice that at $x = x_0$, $L_0(x) = 1$ and $L_1(x) = 0$, similarly at $x = x_1$, $L_1(x) = 1$ and $L_0(x) = 0$, so we have $p_1 = f_0, x = x_0$ and $p_1 = f_1, x = x_1$. This leads to the linear Lagrange polynomial $p_1(x) = \frac{x - x_1}{x_0 - x_1} f_0 + \frac{x - x_0}{x_1 - x_0} f_1$, this is an example of linear interpolation. Quadratic interpolation would require $(x_0, f_0),(x_1, f_1),(x_2, f_2)$, this leads us to the second degree Lagrange polynomial $p_2(x) = L_0(x) f_0 + L_1(x) f_1 + L_2(x) f_2$,

where $L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}$, $L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}$, $L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}$.

The general Lagrange interpolation polynomial is $p_n(x) = \sum_0^n L_k(x) f_k$. It can be shown that $\varepsilon_n(x)$ the error in approximating $f(x)$ by $p_n(x)$ is given by $(x - x_0)(x - x_1)...(x - x_n)\frac{f^{(n+1)}(t)}{(n+1)!}$, where $x_0 \le t \le x_n$. We might argue that given $\varepsilon_n(x) = (x - x_0)(x - x_1)...(x - x_n)\frac{f^{(n+1)}(t)}{(n+1)!}$, then as n becomes large $\varepsilon_n(x)$ becomes small, i.e. the greater the degree of $p_n(x)$ the better the interpolation. Sadly this is not the case in general, there are functions f for which $p_n(x)$ exhibits large oscillations between the nodes, this is an example of Runge's phenomenon (Runge 1901).

4.1.0 Cubic Splines

In trying to approximate some function $f(x)$ by a single a single polynomial it is not uncommon to encounter problems of numerical stability ($p_n(x)$ exhibits large oscillations between the nodes). Splines offer a way of approximating $f(x)$ that can to a reasonable extent avoid problems of numerical instability. Spline interpolation can be defined as piecewise polynomial interpolation. If $f(x)$ is defined on the interval [a,b], then the interval [a,b] is split so that $a = x_0 < x_1 < x_2 < .... < x_n = b$. It can be seen that each subinterval $[x_j, x_{j+1}]$ has a common endpoint, these endpoints are called nodes, in most statistical literature nodes are referred to as knots, I shall follow suit and use the term knot throughout in my discussion of splines. A polynomial $g(x)$ is required such that $f(x_0) = g(x_0),..., f(x_n) = g(x_n)$, also it is required that at the knots $g(x)$ can be differentiated several times, such a $g(x)$ is called a spline. I shall concentrate on cubic splines, a cubic spline $g(x)$ defined on [a,b] is a continuous function and has continuous first and second derivatives, (continuous in [a,b] and all subintervals of [a,b]). Also for each subinterval of [a,b] $g(x)$ is a polynomial of not more than degree 3.

Now by definition $g(x)$ is such that for each subinterval in [a,b], $g(x)$ must be given by $p_j(x)$ where $p_j(x_j) = f(x_j)$, $p_j(x_{j+1}) = f(x_{j+1})$ and $p'_j(x_j) = k_j$, $p'_j(x_{j+1}) = k_{j+1}$. The degree of $p_j(x)$ must not be greater than 3.

It can be seen that by replacing x by x_j and x_{j+1} in $p_j(x)$, where $p_j(x)$ is given by

$$p_j(x) = f(x_j)c_j^2(x-x_{j+1})^2[1+2c_j(x-x_j)] + f(x_{j+1})c_j^2(x-x_j)^2[1-2c_j(x-x_{j+1})]$$
$$+ k_j c_j^2(x-x_j)(x-x_{j+1})^2 + k_{j+1}c_j^2(x-x_j)^2(x-x_{j+1}) \quad \textbf{(1)}$$

and $c_j = \dfrac{1}{x_{j+1} - x_j}$, results are obtained that satisfy the definition for a cubic spline.

Taking the second derivative to get

$p_j''(x_j) = -6c_j^2 f(x_j) + 6c_j^2 f(x_{j+1}) - 4c_j k_j - 2c_j k_{j+1}$ **(2)**

$p_j''(x_{j+1}) = 6c_j^2 f(x_j) - 6c_j^2 f(x_{j+1}) + 2c_j k_j + 4c_j k_{j+1}$ **(3)**

From the fact that $g(x)$ has continuous second derivatives

$p_{j-1}''(x_j) = p_j''(x_j)$ for $j = 1,...,n-1$ **(4)**

Using $j-1$ in the expressions for $p_j''(x_j)$ and $p_j''(x_{j+1})$, the following result is obtained

$c_{j-1}k_{j-1} + 2(c_{j-1} + c_j)k_j + c_j k_{j+1} = 3(c_{j-1}^2(f(x_j) - f(x_{j-1})) + c_j^2(f(x_{j+1}) - f(x_j)))$ **(5)**

The above result is a system of $n-1$ equations; the system has the unique solution $k_1,...,k_{n-1}$, note that $k_1,...,k_{n-1}$ is $g'(x)$ at the knots. Assuming that the knots are equally spaced, say by a distance h, and writing $x_0, x_1, x_2, ..., x_n$ as $x_0, x_1 = x_0 + h, x_2 = x_0 + 2h, ..., x_n = x_0 + nh$. Also $c_j = \dfrac{1}{x_{j+1} - x_j} = \dfrac{1}{h}$.

Hence

$c_{j-1}k_{j-1} + 2(c_{j-1} + c_j)k_j + c_j k_{j+1} = 3(c_{j-1}^2(f(x_j) - f(x_{j-1})) + c_j^2(f(x_{j+1}) - f(x_j)))$, can now be written as $k_{j-1} + 4k_j + k_{j+1} = \dfrac{3}{h}(f_{j+1} - f_{j-1})$ for $j = 1,...,n-1$ **(6)**

Writing $p_j(x) = a_{j0} + a_{j1}(x - x_j) + a_{j2}(x - x_j)^2 + a_{j3}(x - x_j)^3$ and then by looking at the Taylor series for $p_j(x)$ to get

$a_{j0} = p(x_j) = f_j$

$$a_{j1} = p_j'(x_j) = k_j$$

$$a_{j2} = \frac{1}{2} p_j''(x_j) = \frac{3}{h^2}(f_{j+1} - f_j) - \frac{1}{h}(k_{j+1} + 2k_j)$$

$$a_{j3} = \frac{1}{6} p_j'''(x_j) = \frac{2}{h^3}(f_j - f_{j+1}) + \frac{1}{h^2}(k_{j+1} + k_j)$$

Combining the results directly above with **(6)** allows numerical values for the coefficients of $p_j(x)$ to be determined and hence $g(x)$. In Box 2 some further useful properties of splines are discussed.

Box 2 Splines and Elastic Energy

Splines possess an extremely interesting and useful property. For the spline $g(x)$, $g'(a) = f'(a)$ and $g'(b) = f'(b)$. Now using integration by parts

$$\int_a^b g''(x)(f''(x) - g''(x))dx = \int_a^b u\frac{dv}{dx}dx = [uv]_a^b - \int_a^b v\frac{du}{dx}dx$$

$$= [g'''(x)(f'(x) - g'(x))]_a^b - \int_a^b g'''(x)(f'(x) - g'(x))dx = 0$$

Therefore $\int_a^b g''(x) f''(x)dx = \int_a^b (g''(x))^2 dx$ **(8)**

Now considering $\int_a^b [f''(x) - g''(x)]^2 dx$,

$$\int_a^b [f''(x) - g''(x)]^2 dx = \int_a^b f''(x)^2 dx - 2\int_a^b f''(x)g''(x)dx + \int_a^b g''(x)^2 dx$$, using **(8)** to get

$$\int_a^b [f''(x) - g''(x)]^2 dx = \int_a^b f''(x)^2 dx - \int_a^b g''(x)^2 dx$$ **(9)**

The right hand side of **(9)** is ≥ 0, therefore

$$\int_a^b f''(x)^2 dx \geq \int_a^b g''(x)^2 dx$$ **(10)**

I mentioned earlier that a spline as used by an Engineer or Draughtsman is a thin flexible rod. For $g(x)$, $g''(x)$ is an approximation of the curvature of $g(x)$. Treating $g(x)$ as a thin beam or rod we can say that the curvature $g''(x)$ of $g(x)$ is proportional the bending moment of the rod, also $\int_a^b g''(x)^2\,dx$ is proportional to elastic energy stored in the beam (Horn K.P. 1983). If the conditions $g''(a) = 0$ and $g''(b) = 0$ are imposed on a cubic spline, then we have what is known as a natural or restricted cubic spline. The natural spline possesses the property $\int_a^b g''(x)^2\,dx$ is a minimum. When $f(x)$ is approximated using the natural spline $g(x)$, the approximation is one that minimises elastic energy.

4.2.0 Cubic Splines in a Statistical Context

I shall now consider the use of splines in statistics. I shall make recourse to the paper by Wegman and Wright (Wegman & Wright 1983). The background material I have looked at so far concerning splines is what one would find in any useful textbook on Engineering Mathematics, I have not addressed the use of splines in statistical work. The data used in an engineering application of splines is different from the data that might be used in a biostatistical application of splines. Engineering data would tend to be less noisy, Wegman and Wright (Wegman & Wright 1983) state:

"More to the point, it is desirable in a statistical framework to create a type of spline that could pass near, in some sense, to the data but not be constrained to interpolate exactly"

Wegman and Wright point out that in a statistical context fitting a spline goes beyond solving a linear system of equations, we have to consider a 'genuine optimization routine'. Wegman and Wright identify three ways of fitting smoothing splines, viz

penalised least squares, 100 percent confidence intervals and regression splines. I make extensive use of Wegman and Wright's paper (Wegman & Wright 1983). I shall now examine in some detail the three methods as described by Wegman and Wright.

4.2.1 Penalised Least Squares

Using the notation in (Wegman & Wright 1983) for penalised least squares consider the solution to the following optimisation problem

Minimise

$$\sum_{1}^{n}(f(x_j)-y_j)^2+\lambda\int_{0}^{1}(Lf(x))^2\,dx, \text{ subject to } f\in W_m \quad \textbf{(11)}$$

It is assumed that $0<x_1<x_2<\ldots<x_n<1$ and $\lambda>0$ is a fixed parameter, **(11)** is what is known as an objective function. The set of functions f on [0,1] such that $D^j f$, $j\le m-1$ is absolutely continuous and $D^m f$ is in L_2 is denoted by W_m (see note).It can be seen that the integral that appears in **(11)** is similar to $\int_a^b g''(x)^2\,dx$, L is a differential operator, $L=D^m$, where D denotes differentiation, so with $L=D^2$, $Lf(x)$ is equivalent to $\frac{d^2f(x)}{dx^2}$. The term $\lambda\int_0^1(Lf(x))^2\,dx$ is known as a penalty term, it penalises lack of smoothness. I need to introduce the idea of smoothing, when I smooth data I am attempting to fit a curve to the data that picks up important general features, but leaves out fine grained local detail i.e. leaves out the noise. If λ is allowed to get very close to 0, then there is no smoothing, if λ is allowed to become extremely large, in fact let $\lambda\to\infty$, then I have infinite smoothing. As $\lambda\to 0$ then $f(x)$ becomes an interpolating spline, as $\lambda\to\infty$, then $f(x)$ becomes a least squares

estimate. Informally I could describe a smoothing spline as a way of fitting a curve to a dataset with the aim of striking a balance between the interpolation spline which will fit the data to a very high degree and the least squares approach which may not. It is important to distinguish between an interpolation spline and a smoothing spline, the interpolation spline would be the thing to use if I were interested in mathematically describing the shape of curved component in engineering, for example the curve of a wheel arch on a car. In the context of statistical modelling I might argue that the smoothing spline would be an appropriate tool, as in this case we are concerned with general overall patterns and relationships, and not with fine grained detail. I could express these points in terms of over-fitting and under-fitting, the interpolation spline will over-fit, the least squares estimate may lead to under-fitting. What can I say about the smoothing spline in regard to over-fitting and under-fitting? As Wegman and Wright point out the choice of λ is of paramount importance, as the sample size increases then λ should be decreased. Wahba and Wold (Wahba & Wold 1975) develop a method for selecting λ using cross-validation. Wahba and Wold use the following criteria to select λ :

Using Wahba and Wold's notation minimise $E\left[\frac{1}{2n}\sum_{j=1}^{2n}\left(g_{n,\lambda}(x_j)-g(x_j)\right)^2\right]$, i.e. minimise the average mean square error. Note $g_{n,\lambda}(x_j)$ is a spline $g(x_j)$ the observed data. The quantity $\left[\frac{1}{2n}\sum_{j=1}^{2n}\left(g_{n,\lambda}(x_j)-g(x_j)\right)^2\right]$ regarded as a function of λ is known as the cross-validation function, by introducing $w(\lambda)$ (a weighting function) into the expression for the cross-validation function, the generalised cross-validation function is obtained, i.e.

$$\left[\frac{1}{2n}\sum_{j=1}^{2n}\left(g_{n,\lambda}(x_j)-g(x_j)\right)^2 w(\lambda)\right].$$

It can be shown that this function can be represented using matrices, the estimate of λ obtained from the generalised cross-validation function is the best one to use in the penalised least squares method.

4.2.2 100 Percent Confidence Interval Method

The second method for fitting smoothing splines discussed by Wegman and Wright is 100 percent confidence intervals. In (Wegman & Wright 1983) an interpolating spline is considered as the solution to an optimisation problem. Using the notation in (Wegman & Wright 1983) the interpolating spline $s(x)$ is the solution to:

Minimise $\int_{-\infty}^{\infty}(L(f(x))^2\,dx$, subject to $D^j f \in L_2(-\infty,\infty)$, $j=0,1,...,m$ and $f(x_i)=y_i$, $i=1,2,...,n$. **(12)** .

L_2 is a set of measurable integrable square functions. *(<u>Note</u> L_2 same as L^2 , i.e. Lesbague space. Square integrable means $\int|f|^2$ over interval (a,b) is finite)*

Here $\int_{-\infty}^{\infty}(L(f(x))^2\,dx$ is the objective function. The interpolation spline $s(x)$ is a polynomial of degree $2m-1$ .It was seen that for penalised least squares the objective function contained a least squares term $\sum_{1}^{n}(f(x_j)-y_j)^2$, for the 100 percent confidence interval method the objective function is the same as for the penalised least squares case but the interpolating constraints are relaxed. For the 100 percent confidence interval method according to (Wegman & Wright 1983) I have the optimisation problem:

Minimise $\int_{-\infty}^{\infty}(L(f(x))^2\,dx$, subject to $f \in W_m, \alpha_i \le f(t_i) \le \beta_i$, $i=1,2,...,n$

From a statistical point of view the 100 percent confidence interval method can be understood in terms of the model $y_i = f(x_i) + \varepsilon_i$, $i = 1,2,...,n$. Assuming that ε_i is i..i.d on $[-e_1, e_2]$, then because $\varepsilon_i > -e_1$, $y_i + e_1 > y_i - \varepsilon_i = f(x_i)$. Because $\varepsilon_i < e_2$, $y_i - e_2 < y_i - \varepsilon_i = f(x_i)$, so $(y_i - e_2, y_i + e_1)$ is a 100 percent confidence interval. As Wegman and Write point out the 100 percent confidence interval method is an example of the Generalized Hermite-Birkoff interpolation problem (Schoenberg 1966).

4.2.3 Regression Splines

In (Wegman & Wright 1983) the penalised least squares and 100 percent confidence interval methods for fitting smoothing splines are presented as optimization problems, I want to minimise curvature. Regression splines can be regarded in the manner that I first introduced the idea of a spline, a continuous piecewise polynomial of degree m. Regression splines require that I determine several free parameters. I do not have assume that the knots are co-incident with the $x's$, I can choose the number and position of the knots. I can of course choose the degree of the spline. Also I can determine the free coefficients in the spline, there are $m + N + 1$ free coefficients, there are continuity conditions placed on the first $m - 1$ derivatives of the spline, the free coefficients are those remaining after these conditions have been met. Using the notation in (Wegman & Wright 1983) consider the model

$y_i = s_\Delta(x_i) + \varepsilon_i$, $i = 1,2,...,n$. **(13)**

In $s_\Delta(x)$ the symbol Δ denotes a mesh of knots, $\Delta = \{\zeta_1 < \zeta_2 <,... < \zeta_n\}$, where ζ_i is a knot. With N knots, and $N + 1$ polynomial segments of degree m ,(**13**) can be written as $y_i = \sum_{j=0}^{m} \beta_{0j} x_i^j + \sum_{k=1}^{N} \sum_{j=0}^{m} \beta_{kj} (x_i - \zeta_k)_+^j + \varepsilon_i$ **(14)**

Note that in (**14**) the term $(x_i - \zeta_k)_+^j$ is written using Heaviside notation, that is $u(t-a)_+ = u(t-a)$ if $u(t-a) > 0$ and $u(t-a)_+ = 0$ if $u(t-a) \leq 0$.

The big advantage of **(14)** is that I can use ordinary least squares regression to obtain estimates for the coefficients β_{kj}. I said earlier that the knots do not have to be co-incident with the $x's$, Wegman and Wright draw attention to Wold's (Wold 1974) recommendation that knots should be located at data points. Wold (Wold 1974) also recommends that I use as few knots as possible, the more knots that are used the more complex the model, i.e. I have to estimate more parameters. Also a large number of knots may lead to over-fitting. I must exercise caution when choosing the location of the knots, in selecting two adjacent knots I have in effect defined the interval $[\zeta_i, \zeta_{i+1}]$, it might be that within this interval there are points for which the curve passing through the points (x, y) has a minimum or a maximum, or has a point of inflexion. If I wish to use cubic splines this is not a problem provided there are not multiple maximum and minimum points, and there are not multiple points of inflexion. Wold (Wold 1974) notes that if this is the case then we could not employ a cubic spline. A cubic polynomial can have both a maximum and a minimum, and a single point of inflexion, but not multiple maximum and minimum points, and not multiple points of inflexion. According to Wold (Wold 1974), maximum and minimum points should be located at the centre of the interval. Points of inflexion should be located close to the knots. A common choice for m in (**14**) is 3, giving a cubic spline. The cubic spline is popular because it allows researchers to tackle a good range of data sets where a polynomial model is appropriate, the cubic spline avoids the overheads for splines of larger degree. Harrell (Herndon & Harrell 1990), (Harrell *et al.* 1996) advocates the use of cubic splines, specifically the restricted cubic spline. In (Herndon & Harrell 1990) the main focus is on the use of the restricted cubic spline

in connection to the hazard function, Harrell finds that the restricted cubic spline can be used to model data where the hazard function may be one of several different shapes. In (Herndon & Harrell 1990) data from various distributions where considered. Earlier it was said that if the conditions $g''(a) = 0$ and $g''(b) = 0$ are imposed on a cubic spline, a restricted cubic spline is obtained. When trying to model survival data the researcher should be aware that the cubic spline may present problems. Stone and Koo (Stone & Koo 1986) have found that for points beyond the first and last knots the cubic spline may exhibit strange behaviour. The restricted cubic spline does not exhibit strange behaviour at points beyond the first and last knots .The restricted cubic spline is linear at points close to the first and last knots.

4.2.4 Splines applied to the CARE-HF data

The literature on the use of splines in statistics is considerable and large portion is of a high level of mathematical sophistication. I have confined myself to a discussion of some of the basic points. If I want to follow the advice of authors and researchers such a Harrell and adopt the use of splines in modelling how easily is this accomplished? Cubic splines have been implemented in a number of statistical software packages. For SAS the RCS macro (Heinzl & Kaider 1997), (Heinzl & Kaider 2006) is available, for GNU R and S-Plus Harrell's Design (Design Library Harrell Frank E. 2009b) package provides the restricted cubic spline in a form which is easily used in a Cox Proportional Hazards model.

I shall now look at a simple example of using the RCS macro to fit a cubic spline to the CARE-HF data (Richardson *et al.* 2007). The aim of this example is to demonstrate basic usage of the RCS macro and to illustrate a simple and practical approach to the issue of functional form for a model. In the this example I shall fit a Cox Proportional Hazards model with systolic blood pressure and CRT as

independent variables, however I shall include a cubic spline representation of systolic blood pressure in the model. The following SAS code is an example of how I use the RCS macro to fit a cubic spline:

```
%RCS(
  TITLE=%STR(CAREHF),
  DATA=LATESTEX,DIRDATA=%STR(C:\Documents and
Settings\richarmz.ADF.000\Desktop\prog_card_dat\),
  PROGRAM=%STR(C:\Documents and
Settings\richarmz.ADF.000\Desktop\prog_card_dat\rcs\sbpspline.sas),
  TIME=futime,STATUS=primary,
  COV1=supsys,WHAT1=0,KNOTS1=105 117 130 165,
  COV2=treat
  );
```

The reader is directed to (Heinzl & Kaider 1997) for an explanation of the RCS code, however it might be helpful to comment here on the above code. The line beginning with the keyword DATA is where I specify the name and location of a SAS dataset, in this example the dataset is called LATESTEX. The line beginning with the keyword PROGRAM allows me to specify the name and location of the SAS program sbpspline.sas. TIME and STATUS refer to survival time and censoring respectively. On the line beginning COV1 I specify supsys (systolic blood pressure), if set to 1 WHAT1 allows modelling of time by covariate interaction with the cubic spline. The knots for the cubic spline are specified using KNOTS1, I have knots at 105, 117, 130, 165. COV2 specifies that the next independent variable in the model is treat (CRT). NB the above code will not produce any output in terms of analysis. The fitting of the cubic spline is performed by running the SAS program sbpspline.sas, this program calls PROC PHREG, PROC IML and PROC GPLOT. On running sbpspline.sas I obtain output from PHREG and GPLOT . Below is an extract of the output from PHREG.

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Supsys	1	-0.01492	0.00697	4.5829	0.0323	0.985	0.972	0.999
__1_1	1	0.0000110	0.0000212	0.2678	0.6048	1.000	1.000	1.000
__1_2	1	-0.0000172	0.0000532	0.1047	0.7462	1.000	1.000	1.000
treat	1	-0.47246	0.10505	20.2258	<.0001	0.623	0.507	0.766

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
EFFECT1	9.7550	3	0.0208
NONLIN1	2.0941	2	0.351

It can be seen that systolic blood pressure (supsys) is a significant predictor of time to death or unplanned hospitalisation as is whether or not a patient has received cardiac resynchronisation (treat). In the parameter column of the output __1_1 and __1_2 refer to the cubic spline representation of systolic blood pressure, neither are statistically significant. From this I would conclude that a cubic spline representation of systolic blood pressure does not represent an improvement in functional form over the assumed linear form, this is reflected in the linear hypotheses testing results.

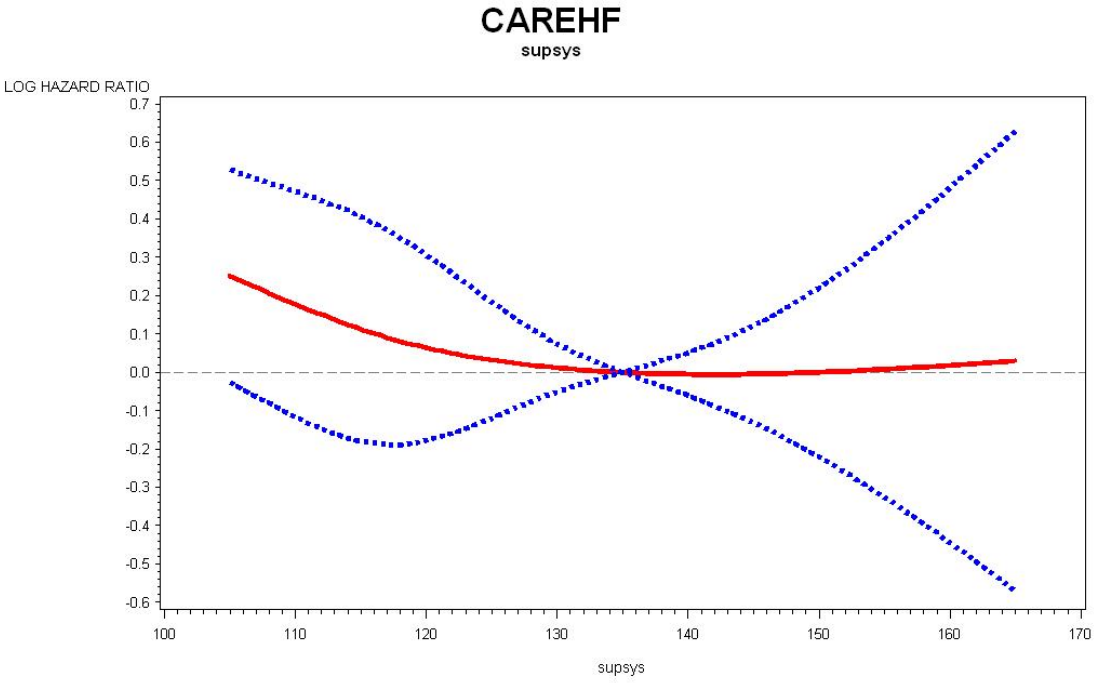


Figure 4.2 The restricted cubic spline (the red curve) approximation for the log hazard ratio as a function of systolic blood pressure.

The figure above shows the restricted cubic spline (the red curve) approximation for the log hazard ratio as a function of systolic blood pressure (supsys). The blue dotted curves represent the confidence curves, at first glance one might think that there are two curves that cross, this is not the case. At the middle of the spline the two confidence curves are very close together. Using the notation and derivation from (Heinzl & Kaider 1997) the restricted cubic spline is given by

$C(u) = \beta_0 + \beta_1 u + \sum_{j=1}^{k-2} \theta_j C_j(u)$, where k is the number of knots, let the knots be

$t_1, t_2, ...t_k$. Also $C_j(u) = (u - t_j)_+^3 - \frac{(u - t_{k-1})_+^3 (t_k - t_j)}{(t_k - t_{k-1})} + \frac{(u - t_k)_+^3 (t_{k-1} - t_j)}{(t_k - t_{k-1})}$.

In the output from PHREG__1_1 and __1_2 refer to $C_1(u)$ and $C_2(u)$ respectively, estimates for θ_1 and θ_2 are 0.0000110 and -0.0000172.

Again using the notation and derivation in (Heinzl & Kaider 1997) for a fixed value u_0 the estimated cubic function $\hat{C}(u_0)$ can be written as $\hat{C}(u_0) = \hat{\beta}' U_0$, where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\theta}_1, ..., \hat{\theta}_{k-2})'$ and $U_0 = (1, u_0, C_1(u_0), ..., C_{k-2}(u_0))^t$. If V is the sample covariance matrix for $\hat{\beta}$ then a $1-\alpha$ confidence interval for $\hat{C}(u_0)$ is given by

$\hat{\beta}' U_0 \pm (\gamma U_0' V U_0)^{\frac{1}{2}}$, $\gamma = \chi^2_{p,1-\alpha}$ is the $1-\alpha$ quantile of χ^2 with p degrees of freedom (Heinzl & Kaider 1997). To understand why the two confidence curves are very close together at the middle of the spline, note that $\hat{\beta}' U_0 - (\gamma U_0' V U_0)^{\frac{1}{2}}$ and $\hat{\beta}' U_0 + (\gamma U_0' V U_0)^{\frac{1}{2}}$ will increase in size as u_0 moves further from the mean, i.e. the distance between the confidence curves increases as u_0 moves further from the mean.

Further material on the use of splines in statistics is to be found in (Smith 1979) and (Poirier 1979). I now consider fractional polynomials.

4.3.0 Fractional Polynomials

The cubic spline is one example of using polynomials to model data. Another approach is that of fractional polynomials, see (Royston Patrick *et al.* 1999), (Royston & Altman 1994), (Stocken D.D. *et al.* 2008), (Royston & Sauerbrei 2004) and (Meier-Hirmer *et al.* 2003). In epidemiological and biostatistical applications continuous variables such as age are often split into groups to form a new categorical variable. This makes analysis easy perform and interpret; however in doing this the researcher may encounter problems. If I have not pre-specified how I intend to form the groups, that is the location of the cut-points or group boundaries, I can end up with highly 'data driven' results. Also in moving from continuous to categorical data I introduce 'jumps' when a group boundary or cut point is crossed, for example if I where modelling the probability of some event occurring as a function of age, the probability of the event occurring will jump, perhaps quite substantially when a cut point is crossed. Is this a realistic model of the situation? Altman and Royston (Altman & Royston 2006) state that dichotomising variables leads to loss of information, reduced statistical power and an increased risk of false positive results. In view of this, there is an argument for preserving continuous data. As shown earlier cubic splines can be used to model the relationship between the dependent variable Y and the independent variable X when the relationship is not a simple linear one. The fractional polynomial developed by Royston and Altman (Royston & Altman 1994) allows the researcher to consider a number of possible functional forms for the relationship between Y and X . In (Royston & Altman 1994) the fractional polynomial is defined as follows

$\phi_m(X;\xi,p) = \xi_0 + \sum_{j=1}^{m} \xi_j X^{(p_j)}$, where $X > 0$ and $p = (p_1,...,p_m)$ is a vector of powers with $p_1 < ... < p_m$ and $\xi = (\xi_0,\xi_1,...,\xi_m)$ a vector of coefficients, both p and ξ are real valued. Also $X^{(p_j)} = X^{p_j}$ if $p_j \neq 0$, $X^{(p_j)} = \log_e X$ if $p_j = 0$; the Box-Tidwell Transformation.

Royston and Altman give what they say is their full and most concise definition as follows

$\phi_m(X;\xi,p) = \sum_{j=0}^{m} \xi_j H_j X^{(p_j)}$, where for $j = 1,...,m$ and $H_j(X) = X^{(p_j)}$ if $p_j \neq p_{j-1}$, $H_j(X) = H_{j-1}(\log_e X)$ if $p_j = p_{j-1}$. In (Royston & Altman 1994) the authors state that for given values of m and p the fractional polynomial given in the form above can be regarded for the purpose of model development as a linear predictor. The best values for m and p need to be determined, in (Royston & Altman 1994) the authors suggest that for most practical situations $p = \{-2,-1,-0.5,0,0.5,1,2,...,\max(3,m)\}$ is adequate. The degree of the fractional polynomial m is determined on an informal basis *a priori* or until no appreciable improvement in model fit is observed. It can be seen that fractional polynomials obtained using $p = \{-2,-1,-0.5,0,0.5,1,2,...,\max(3,m)\}$ contains the straight line case, the natural log, the square root. The fractional polynomial is flexible in the sense that it allows me to fit many of the 'standard' models. I could view the fractional polynomial as a generalised method for applying transformations. The fractional polynomial allows me to produce a model with a sensible functional form. In regard to model fit Royston and Altman assume that maximum likelihood is used. Based on a given m the best vector of powers $\tilde{p}$ is the one from the model with the greatest likelihood or the

smallest deviance D. In (Royston & Altman 1994) the authors use the quantity $D(m, p) - D(m, \hat{p})$ which is distributed (asymptotically) χ^2 with m degrees of freedom , $\hat{p}$ is the full maximum likelihood estimate of p. This quantity may be used to assess the adequacy of a conventional polynomial versus a fractional polynomial of the same degree. Another quantity which is defined in (Royston & Altman 1994) as the gain $G = G(m, p) = D(1,1) - D(m, p)$ uses the deviance of the straight line model $D(1,1)$ as a reference against which to compare other models. Unlike the deviance a large value for the gain is an indication of a better fit. In the definitions so far of a fractional polynomial we are dealing with a single independent variable, it is possible to extend the definition of a fractional polynomial to include several independent variables

Multivariable fractional polynomials are implemented in SAS via the %mfp8 macro (MFP 2009).The %mfp8 macro has been ported to GNU R as the mfp library . I shall now look at an example using the mfp library in GNU R to fit fractional polynomials to the CARE-HF data (Richardson *et al.* 2007). The following R code demonstrates basic usage of the mfp library:

```
setwd("C:/Documents and
Settings/richarmz.ADF.000/Desktop/phd_chapters")
dd<-read.Table(file="latest_ex2.csv",header=T,sep=",")
attach(dd)
library(mfp)
f<-
mfp(Surv(futime,primary)~fp(Roche)+fp(mitral_r)+fp(Supsys)+Ischaemic+
treat,select=0.05,verbose=TRUE,family=cox,data=dd)
```

here I am fitting a Cox Proportional Hazards model which incorporates fractional polynomials for N-terminal pro-brain natriuretic peptide (Roche), Mitral regurgitation (mitral_r), and Systolic blood pressure (Supsys). An extract of the GNU R output is shown below:

```
          df.initial select alpha df.final power1 power2
mitral_r           4   0.05  0.05        4     -2      2
Ischaemic          1   0.05  0.05        1      1      .
treat              1   0.05  0.05        1      1      .
Roche              4   0.05  0.05        2      0      .
Supsys             4   0.05  0.05        0      .      .
```

```
Transformations of covariates:

                                          formula
Roche                          log((Roche/10000))
mitral_r  I((mitral_r/10)^-2)+I((mitral_r/10)^2)
Supsys                                      <NA>
Ischaemic                               Ischaemic
treat                                       treat
```

The mfp function selects the best fitting fractional polynomial. The natural log transformation of N-terminal pro-brain natriuretic peptide (Roche) has been selected. This result is in accord with the findings in (Richardson *et al.* 2007), i.e. on comparing the AIC for two Cox Proportional Hazards models of the form $X + (X * CRT) + CRT$ and $\log_e X + (\log_e X * CRT) + CRT$ is was found that for N-terminal pro-brain natriuretic peptide the model that used the natural logarithm transform result in a smaller AIC. For mitral regurgitation (mitral_r) a fractional

polynomial of the form $c_1\left(\frac{x}{10}\right)^{-2}+c_2\left(\frac{x}{10}\right)^{2}$ has been selected. The coefficients c_1 and c_2 can be obtained in GNU R , they are -30.6 and 0.000169 respectively. In (Richardson *et al.* 2007) the logarithmic transformation applied to mitral regurgitation was found to improve model fit. The transformation selected on the basis of a statistically significant difference in the AICs for models of the form $X+(X*CRT)+CRT$ and $\log_e X+(\log_e X*CRT)+CRT$, may well be different from those obtained by using mfp in the way just demonstrated. For Systolic blood pressure (supsys) has been omitted from the 'final' model, in (Richardson *et al.* 2007) systolic blood pressure was included because the interaction term (systolic blood pressure*CRT) was found to be statistically significant. Dealing with interaction terms in mfp involves setting up a new variable, I cannot use a term such as (supsys*CRT) in mfp, i.e. I cannot explicitly write an interaction term. Instead I would create a new variable, for example supt=(supsys*CRT). After doing this it is possible to include the interaction term using in the following code in GNU R:

```
f2<-
mfp(Surv(futime,primary)~fp(Supsys)+supt+treat,select=0.05,verbose=TR
UE,family=cox,data=dd)
```

If the above code is run then the results are in agreement with those found in (Richardson *et al.* 2007), systolic blood pressure (supsys) is left un-transformed, also the hazard ratios and p-values for systolic blood pressure, CRT (treat) and the interaction term are as reported in Table 2 of (Richardson *et al.* 2007) (these models where produced using PHREG in SAS). Similarly for mitral regurgitation if a new

variable is set up for the interaction with CRT, then mfp reports that the best fractional polynomial for mitral regurgitation is the natural logarithm.

A question can be raised in regard to the attempt at fitting a fractional polynomial to supt=(supsys*CRT) i.e. including `fp(supt)` in the model statement above. Is this valid or would it be better to use another method of fitting the model? If a fractional polynomial is fitted for the interaction term the following output is obtained:

```
       df.initial select alpha df.final power1 power2
Supsys          4   0.05  0.05        1      1      .
treat           1   0.05  0.05        1      1      .
supt            4   0.05  0.05        2      3      .
```

```
Transformations of covariates:
                   formula
Supsys   I((Supsys/100)^1)
supt   I(((supt+1)/100)^3)
treat                treat
```

```
               coef exp(coef)  se(coef)      z        p
Supsys.1 -1.466e-02    0.9854 3.950e-03 -3.711 2.06e-04
treat.1  -1.118e+00    0.3270 2.513e-01 -4.449 8.63e-06
supt.1    3.759e-07    1.0000 1.332e-07  2.822 4.77e-03
```

Here the interaction term supt itself has undergone a non-linear transformation, I cannot interpret the transformed interaction term in an obvious way , the main effect Supsys is untransformed whereas supt is now a cubic term. The p-values for systolic blood pressure, CRT (treat) and the interaction term are smaller than those reported in Table 2 of (Richardson *et al.* 2007). It would be better to establish the fractional polynomial for the main effect first and then fit a model that uses the transformed (or un-transformed) variable for both the main effect and the interaction term. If in the example models below Z is a binary variable and $f()$ is some transformation then when using mfp model 4 produces the same results as model 2 using PHREG, whereas model 3 using mfp produces different results to model 2 using PHREG.

Example Models

1. $X + (X * Z) + Z$
2. $f(X) + (f(X) * Z) + Z$
3. $f(X) + f(X * Z) + Z$
4. $f(X) + (X * Z) + Z$

For instance in GNU R I could use coxph to fit the Cox Proportional Hazards model with the transformed variables obtained from mfp. Ischaemic and treat (CRT) remain un-transformed of course. It should be remembered that when using fractional polynomials the independent variables are assumed to be positive. If the preceding code is run, but this time a fractional polynomial for interventricular mechanical delay is included then mfp produces warnings concerning the failure of the algorithm to converge. By default mfp should shift and scale variables to avoid numerical problems if negative values are present, as is the case for interventricular mechanical delay. I have noted that interventricular mechanical delay is indeed shifted and scaled, yet the warnings from mfp persist, this is the case even if manual shifting and scaling is employed.

4.4.0 Splines versus Fractional Polynomials

Is it better to use splines or fractional polynomials in statistical modelling? Both methods have very appealing aspects. The fractional polynomial is elegant and compact; we can see that the standard transformations are continued with the definition of a fractional polynomial. Does the piecewise nature of the spline afford an advantage over the fractional polynomial? Royston and Altman criticise conventional polynomials as often not providing a particularly good fit. In their view cubic splines are considered to be too computationally intensive, and not amenable to easy interpretation. Also splines are not implemented in standard regression software. Splines do not provide equations that can be easily used for prediction. Royston and Altman made these remarks back in 1994, from a computational perspective things have moved on, the software is now available and fast processors now make it quite feasible to fit cubic splines routinely. Royston and Altman also suggest that the concept of splines is difficult to explain to a ‘non-expert’ user. I take the view that the

GNU R port of the %mfp8 macro produces output that is easier to interpret than that produced in the SAS version. This is a matter of personal taste, but clear reporting of analysis using fractional polynomials is vital, when first encountered fractional polynomials can be somewhat confusing, at first it can be a little difficult to determine what exactly the best polynomial is. In (Royston & Altman 1994) the authors state that fractional polynomials tend not to display the same degree of strange behaviour near endpoints as that of conventional polynomials. This issue is of great interest to the present author, investigation of the numerical stability of cubic splines compared to fractional polynomials would be a useful area for future research. If appropriate a more simple approach such as fitting a quadratic or a cubic term should not be abandoned, this approach avoids the need for additional macros and there is no doubt that the resulting model can be a lot easier to interpret. I would recommend that simple transformations such as these are applied before recourse to more complex methods. Dichotomising continuous variables is widely used in medical and epidemiological applications. Although as discussed earlier good reasons can be supplied to avoid dichotomising continuous variables, however this approach does result in a model that is easier to interpret than one which includes say cubic splines. I would suggest that categorising continuous variables if done sensibly is a perfectly reasonable approach.

I have talked about specifying the functional form for a model, but so far I have not discussed a means of selecting between different models. For example if I wish to establish whether using the natural log transformation has any benefit, I need to compare the model using the transformation with the model without the transformation. If I were to see an improvement in the fit of the model using the

transformation, then I would consider the transformation beneficial. Here I am presented with the problem of model selection. In the next chapter I consider the use of the AIC (Akaike Information Criteria) as a model selection tool.

CHAPTER 5 MODEL FIT, LIKELIHOOD, THE AIC

- The AIC is a penalised log likelihood model selection criterion.
- A modified AIC is required for small samples or where the number of model parameters is large relative to the sample size.
- There are issues with the AIC regarding estimation of the order of the 'true' model. The AIC posits a 'true' model of infinite order.
- The BIC posits a 'true' model of small dimension, the BIC is said to be dimension consistent.
- AIC and variants implemented for mixed models in SAS via GLMMIX, NLMIXED and MIXED.
- AIC and BIC implemented for models with time dependent covariates in SAS via PHREG.
- AIC for frailty models, further investigation may be required

5.0.0 Introduction

In the previous chapter the question of the functional form of a model was discussed. The functional form will have an effect on how well a model fits the data. For example would taking the natural logarithm of one or more of the independent variables or fitting a cubic spline lead to an improvement in the fit of a model. In this chapter the idea of model fit is investigated in more general terms, the idea of selecting the 'best' model is considered. I now review some standard topics in likelihood theory. I shall concentrate on the idea of likelihood and selection criteria based upon the concept of maximising the likelihood. I wish to make it absolutely

clear that all of the mathematical derivations in this chapter are of known results attributable to others, and similar derivations may be found in a number of classic texts. A very comprehensive treatment of likelihood theory can be found in Pawitan's book (Pawitan 2001). The graphical figures in this chapter where produced by myself using simulated data in GNU R (R Foundation for Statistical Computing 2009).

5.1.0 Likelihood

Likelihood plays a central role in statistical modelling. R.A Fisher (Fisher 1932), (Fisher 1934a) and (Fisher 1934b) formulated the idea of likelihood as a middle ground between the Bayesian and frequentist camps. A basic distinction between the Bayesian and frequentist approach can be made with reference to the meaning of statements such as the probability of observing a HEAD with a fair coin is 0.5. The frequentist would insist that the value 0.5 is only meaningful as a long run measure. If the coin where to be tossed a second time the Bayesian would be quite happy to say that that his or her degree of belief that the coin would show a HEAD was 0.5. The frequentist would say that this value is only meaningful in the long run. Note a Bayesian would also accept the idea of a probability being a long run measure. The important point so far as a discussion of likelihood is concerned is that both Bayesians and frequentists make inferences based on probability. With likelihood methods the likelihood function is used to make inferences, inference is not made using 'pure' probability. If I toss a coin 5 times and observe the sequence HEADS, TAILS, HEADS, HEADS, TAILS, then the probability of observing this sequence is $p \times (1-p) \times p \times p \times (1-p)$ or $p^5 + p^3 - 2p^4$, where p is the probability of observing HEADS and *1-p* is the probability of observing TAILS. If X denotes the number of

heads we observe then in the example above $P(X = 3; n = 5) = p^5 + p^3 - 2p^4$.

$P(X = 3; n = 5) = p^5 + p^3 - 2p^4$ is called the likelihood function, denoted by L .What value of p makes the sequence HEADS, TAILS, HEADS, HEADS, TAILS most likely?

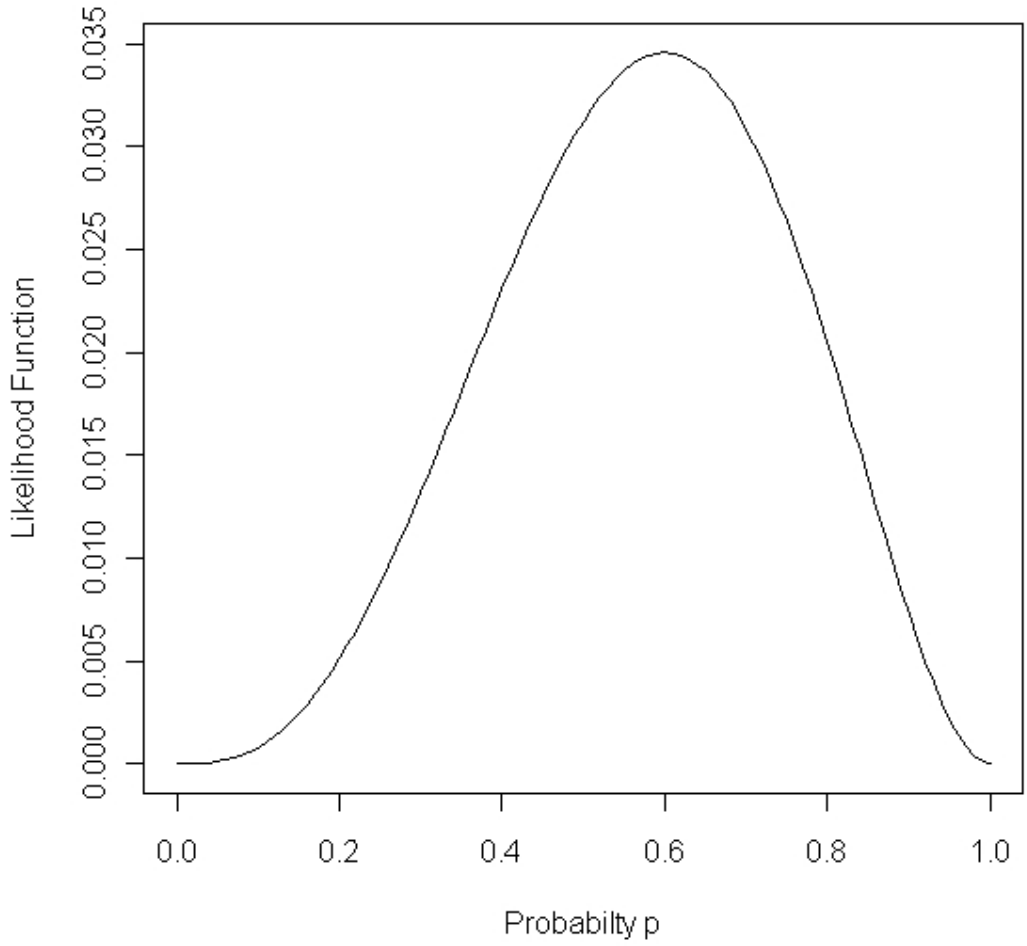


Figure 5.1 Likelihood Function versus Probability

In figure 5.1 a plot of $P(X = 3; n = 5) = p^5 + p^3 - 2p^4$ against p shows that the likelihood is function is a maximum for $p = 0.6$, this value of p for which the likelihood function is a maximum is known as a maximum likelihood estimate (MLE). I know that for a fair coin $p = 0.5$, from figure 5.1 I see that for $p = 0.5$ the

sequence HEADS, TAILS, HEADS, HEADS, TAILS is less likely to be observed. If I were to toss the coin say 1000 times and observed around 500 HEADS , then a plot of the likelihood function against *p* would show that the likelihood function is a maximum for $p = 0.5$. In the coin tossing example above if I believed p to be 0.01, then I obtain a likelihood of 9.8×10^{-7}, if I believed *p* to be 0.6, then I obtain a likelihood of 0.03456. Likelihood can be said to provide a measure of belief. The likelihood principle states that all the information about a sample is contained within the likelihood function. The MLE can also provide evidence to support or contradict our belief, if I believe a coin to be fair (i.e. *p*=0.5) then if for example I obtained a MLE of *p* = 0.89, I have evidence contrary to my belief, if I estimate *p* to be close to 0.5 then I have evidence to support my belief. Population parameters such as *p* are generally denoted by θ, in the following treatment of likelihood theory I will confine myself to the single parameter case, however the type of problem which is the concern of this thesis (fitting a Cox model) requires a multi parameter formulation of likelihood theory. For discrete data we can write $L(\theta) = P(X = x)$. Continuous data presents a problem, I cannot talk about the probability of a continuous variable being exactly equal to a particular value, e.g. $P(X = x)$ is not meaningful. However I can talk about the probability of a continuous variable lying with an interval $(x - \frac{a}{2}, x + \frac{a}{2})$ around *x*. If the interval $(x - \frac{a}{2}, x + \frac{a}{2})$ is small then

$L(\theta) = \int_{x-\frac{a}{2}}^{x+\frac{a}{2}} f(x;\theta)dx$, where $f(x;\theta)$ is the probability density function (p.d.f). It is possible to approximate $L(\theta) = \int_{x-\frac{a}{2}}^{x+\frac{a}{2}} f(x;\theta)dx$ by $af(x;\theta)$, where a is very small, this

approximation is valid only if the data is precise. If I now consider X_1 and X_2, where X_1 and X_2 are identically independently distributed (i.i.d) then

$L_1(\theta) = \int_{x_1-\frac{a}{2}}^{x_1+\frac{a}{2}} f(x;\theta)dx = af(x_1;\theta)$ and $L_2(\theta) = \int_{x_2-\frac{a}{2}}^{x_2+\frac{a}{2}} f(x;\theta)dx = af(x_2;\theta)$ I may combine these likelihoods to give $L(\theta) = L_1(\theta)L_2(\theta) = af(x_1;\theta)af(x_2;\theta)$. For discrete data I have $L(\theta) = P(X_1 = x_1)P(X_2 = x_2)$. For continuous data I notice the presence of the constant *a* in the expressions for $L(\theta)$, the constant *a* can in fact be omitted from the expressions for $L(\theta)$, this can be justified by using the following argument. Consider the model $f(x;\theta)$, (note a p.d.f can be described as a model), further consider the likelihood with different values for θ, θ_1 and θ_2. I wish to compare $L(\theta_1)$ and $L(\theta_2)$, let the likelihood ratio (Note I shall discuss the likelihood ratio in greater detail later in this chapter) $\frac{L(\theta_2)}{L(\theta_1)} = b$, then $L(\theta_1)$ and $L(\theta_2)$ are only meaningful up to a constant multiplier, we have $L(\theta_2) = bL(\theta_1)$, so if I were to consider multiples of $L(\theta_1)$, $aL(\theta_1)$ is only meaningful for *a* up to *b*. In view of this I may write

$L(\theta) = \int_{x-\frac{a}{2}}^{x+\frac{a}{2}} f(x;\theta)dx \approx f(x;\theta)$ and for combined likelihoods $L(\theta) = L_1(\theta)L_2(\theta) = f(x_1;\theta)f(x_2;\theta)$. In general I have

$L(\theta) = \prod_i^n P(X_i = x_i)$ discrete case

$L(\theta) = \prod_i^n f(x_i)$ continuous case

It is mathematically more convenient to work with the natural logarithm of the likelihood function, i.e. $\log_e(L(\theta))$. So I have for the discrete case

$$\sum_i^n \log_e(P(X_i = x_i))$$

And for the continuous case

$$\sum_i^n \log_e(f(x_i))$$

Often interest is focused on obtaining a point estimate of some population parameter, e.g. the sample mean $\bar{x}$ as estimate of the population mean, or s^2 as an estimate of population variance σ^2. The MLE offers another way of obtaining a point estimate, but it is of great importance that attention be paid to the general shape of the likelihood function. Likelihood is a valuable tool in situations where the data may not provide a great deal of information, and where there is a degree of uncertainty. The coin example from earlier represents a situation where I have a small amount of data, I cannot ignore the fact that conclusions I make about this data will be quite uncertain. I wish to maximise $L(\theta)$ or $\log_e L(\theta)$ i.e. I want to find θ such that $\frac{\partial}{\partial\theta}\log_e L(\theta) = 0$. I said earlier that it is important to consider the overall shape of the likelihood function, if for example I have obtained a MLE of θ, $\hat{\theta}$, how certain am I that $\hat{\theta}$ is the 'best' estimate of θ? This question can be answered by looking at the curvature of the likelihood function. θ. If $\theta = \hat{\theta}$ is a solution of $\frac{\partial}{\partial\theta}\log_e L(\theta) = 0$, then $-\frac{\partial^2}{\partial\theta^2}\log_e L(\theta = \hat{\theta}) < 0$, also if $-\frac{\partial^2}{\partial\theta^2}\log_e L(\theta = \hat{\theta})$ is large then $\log_e L(\theta)$ has a tight or sharp peak, this is interpreted as meaning that there is less uncertainty in

regard to my estimate of θ. If $-\frac{\partial^2}{\partial\theta^2}\log_e L(\theta=\hat{\theta})$ is small, then $\log_e L(\theta)$ will not have a sharp peak, this means that there are a number of values of $\hat{\theta}$ that are quite close to the solution of $\frac{\partial}{\partial\theta}\log_e L(\theta)=0$. Put simply I am uncertain as to what numerical value of $\hat{\theta}$ maximises $\log_e L(\theta)$. The quantity $-\frac{\partial^2}{\partial\theta^2}\log_e L(\theta=\hat{\theta})<0$ is known as the observed Fisher information (as the sample size increase then the Fisher information increases). There are many instances when a solution (a closed form solution) of $\frac{\partial}{\partial\theta}\log_e L(\theta)=0$ is not possible, in such cases I am obliged to use numerical methods to obtain an approximate solution. Taking the Taylor series of $\log_e L(\theta)$ about $\hat{\theta}$ I have

$$\log_e L(\theta)\approx\log_e L(\hat{\theta})+(\theta-\hat{\theta})\frac{\partial}{\partial\theta}\log_e L(\hat{\theta})+\frac{(\theta-\hat{\theta})^2}{2}\frac{\partial^2}{\partial\theta^2}\log_e L(\hat{\theta})$$

The above is a quadratic approximation of $\log_e L(\theta)$, in order to make the expression a little more compact denote $\frac{\partial}{\partial\theta}\log_e L(\theta)$ by $S_C(\theta)$ and $-\frac{\partial^2}{\partial\theta^2}\log_e L(\theta)$ by $F_I(\theta)$, then I have

$$\log_e L(\theta)\approx\log_e L(\hat{\theta})+(\theta-\hat{\theta})\,S_C(\hat{\theta})-\frac{(\theta-\hat{\theta})^2}{2}\,F_I(\hat{\theta}).$$

If a quadratic approximation is a good fit for $\log_e L(\theta)$ then $\log_e L(\theta)$ is said to be regular. For regular log likelihood function $\hat{\theta}$ and $F_I(\theta)$ can be used to represent $\log_e L(\theta)$. The following example may help clarify some of the ideas discussed above. Let $x_1, x_2, x_3, ..., x_n$ be a i.i.d sample from a normal distribution with

parameters θ, σ^2. I have $f(x_i;\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}$.

Also $L(\theta) = f(x_1;\theta) f(x_2;\theta) ... f(x_n;\theta)$, hence it is seen that

$\log_e L(\theta) = n \log_e \frac{1}{\sqrt{2\pi}\sigma} - \sum_1^n \frac{(x_i-\theta)^2}{2\sigma^2}$, and $S_C(\theta) = \frac{\sum_1^n (x_i - \theta)}{\sigma^2}$.

The Fisher Information $F_I(\theta) = \frac{n}{\sigma^2}$, also $\text{var}(\hat{\theta}) = \frac{\sigma^2}{n}$, hence $(F_I(\theta))^{-1} = \text{var}(\hat{\theta})$. The connection can be seen between the Fisher Information and the variance of $\hat{\theta}$, i.e. the connection between curvature (measure of uncertainty) of the likelihood function and the variance of $\hat{\theta}$. Fisher Information is of fundamental importance in likelihood theory, in a later section I shall some of the basic ideas in what is known as Information theory. Information theory is highly relevant to the discussion of the Akaike Information Criteria.

5.2.0 Likelihood Ratio

Given a dataset it is possible to fit any number of models, amongst these models some may fit the data quite well, others not so well. A method of comparing these models is required in order that the 'best' one is selected. I can compare two models by examining the likelihood ratio $\frac{L(\mu_a, y)}{L(\mu_b, y)}$ .Note μ_a refers to the full or saturated model m_f and μ_b to some model m_b.

So $\frac{L(\mu_f, y)}{L(\mu_b, y)} \approx 1$ or $\log_e\left(\frac{L(\mu_f, y)}{L(\mu_b, y)}\right) = 0$ suggests that m_b is a good fit.

Let $\varsigma = \log_e\left(\frac{L(\mu_f, y)}{L(\mu_b, y)}\right) = \log_e(L(\mu_f, y)) - \log_e(L(\mu_b, y))$, large values of ς indicate that m_b is a poor fit to the data. The quantity $2\log_e\left(\frac{L(\mu_f, y)}{L(\mu_b, y)}\right)$ is known as the deviance, and is usually denoted by $D(y, \mu)$. It is important to remember that when using the deviance to assess goodness of fit circumstances can easily arise that render the deviance useless as a means of gauging this. If we want to compare two nested models m_1 and m_2 we examine the change in the deviance

$$D(y, \mu_1) - D(y, \mu_2) = 2\log_e\left(\frac{L(\mu_f, y)}{L(\mu_1, y)}\right) - 2\log_e\left(\frac{L(\mu_f, y)}{L(\mu_2, y)}\right) = 2\log_e\left(\frac{L(\mu_2, y)}{L(\mu_1, y)}\right)$$

The deviance follows the χ^2 distribution with $df_1 - df_2$ degrees of freedom.

With nested models I use the deviance to assess whether a term is significant or not, for example I may want to compare the model $Y = c + \beta_1 X_1 + \beta_2 X_2$ with the model $Y = c + \beta_1 X_1$. I might be interested in whether X_2 is significant or not, I look at the change in the deviance due to the inclusion/exclusion of X_2. Note it should be remembered that for each of the models in the above example the deviance is a comparison of the fitted model to the full model.

I must bear in mind that when comparing nested models I am assuming that ϕ the dispersion parameter is equal to 1, if this is not the case then $D(y, \mu_1) - D(y, \mu_2)$ is not meaningful. In situations where $\phi \neq 1$ I use what is known as the scaled deviance $\frac{D(y, \mu_1) - D(y, \mu_2)}{\phi}$.

Some discussion of ϕ is worthwhile, to understand the dispersion parameter I need to consider the Exponential family of distributions.

5.3.0 The Exponential family of distributions

I now review some standard results relating to the Exponential Family of Distributions (Dobson 2002) contains a useful section on this topic. A distribution belongs to the exponential family if it is possible to write $f(x;\theta)$ in the form $u(x)v(\theta)e^{a(x)b(\theta)}$, where u, v, a, b are all known functions. Let $u(x) = e^{g(x)}$ and $v(\theta) = e^{h(\theta)}$, then I may write $f(x;\theta) = e^{g(x)+h(\theta)+a(x)b(\theta)}$. For example consider the Poisson distribution $f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!}$, $f(x;\theta)$ can be written $e^{(x\log_e(\theta)-\theta-\log_e(x!))}$.

By definition $\int_\alpha^\beta f(x;\theta)dx = 1$, then $\frac{d}{d\theta}\int_\alpha^\beta f(x;\theta)dx = 0$ and $\int_\alpha^\beta \frac{d}{d\theta} f(x;\theta)dx = 0$.

Using the same approach I find that $\int_\alpha^\beta \frac{d^2}{d\theta^2} f(x;\theta)dx = 0$.

Now in general $f(x;\theta) = e^{g(x)+h(\theta)+a(x)b(\theta)}$ so $\frac{d}{d\theta} f(x;\theta) = f(x;\theta)(b'(\theta)a(x) + h'(\theta))$.

Using $\int_\alpha^\beta \frac{d}{d\theta} f(x;\theta)dx = 0$, I get $\int_\alpha^\beta f(x;\theta)(b'(\theta)a(x) + h'(\theta))dx = 0$ which can be written as $b'(\theta)E[a(x)] + h'(\theta) = 0$, or $E[a(x)] = \frac{-h'(\theta)}{b'(\theta)}$.

I have $\frac{d^2}{d\theta^2} f(x;\theta) = f(x;\theta)[b''(\theta)a(x) + h''(\theta) + (a(x)b'(\theta) + h'(\theta))^2]$,

$(a(x)b'(\theta) + h'(\theta))^2 = (b'(\theta))^2\left(a(x) + \frac{h'(\theta)}{b'(\theta)}\right)^2 = (b'(\theta))^2(a(x) - E[a(x)])^2$. Also

$\frac{d^2}{d\theta^2} f(x;\theta) = f(x;\theta)[b''(\theta)a(x) + h''(\theta) + (b'(\theta))^2(a(x) - E[a(x)])^2]$, this leads to

$$\int_{\alpha}^{\beta} \frac{d^2}{d\theta^2} f(x;\theta) = b''(\theta)E[a(x)] + h''(\theta) + (b'(\theta))^2 \operatorname{var}[a(x)] = 0 .$$

So $\operatorname{var}[a(x)] = \dfrac{-b''(\theta)E[a(x)] - h''(\theta)}{(b'(\theta))^2}$, but $E[a(x)] = \dfrac{-h'(\theta)}{b'(\theta)}$, so I get

$$\operatorname{var}[a(x)] = \frac{b''(\theta)h'(\theta) - b'(\theta)h''(\theta)}{(b'(\theta))^3} .$$

Returning to the Poisson distribution I have $b = \log_e(\theta)$, b is what is known as a natural parameter. Now $\dfrac{d}{db}E[a(x)] = -h'(\theta)\dfrac{d^2\theta}{db^2}$, $\theta = e^b$, therefore $\dfrac{d}{db}E[a(x)] = -h'(\theta)e^b$ and so I may write $\dfrac{1}{-h'(\theta)}\dfrac{d}{db}E[a(x)] = e^b = \theta$, but from the fact that $\theta = e^b$ I must have $\dfrac{d}{db}E[a(x)] = e^b$, therefore $\dfrac{1}{-h'(\theta)} = 1$. In fact $\dfrac{1}{-h'(\theta)}$ is the dispersion parameter ϕ, for the Poisson distribution I have $\phi = 1$. In general I have $\operatorname{var}(X) = \phi\dfrac{\partial}{\partial b}E[X]$. The dispersion parameter is of great importance in that it allows for a more flexible relationship between the mean and the variance. Certain distributions have limitations as far as statistical modelling is concerned; this is due to the relationship between the mean and the variance. For the binomial distribution I have $\mu = np$ and $\sigma^2 = npq$, I see that the mean and variance are related. When modelling data using Binomial distribution I can encounter the following problem. The data exhibits a larger degree of variability than that assumed from the Binomial distribution. The converse situation can also occur, the data is found to have a smaller degree of variability than that expected from the Binomial distribution. In these situations I have over dispersion and under dispersion. The Exponential Dispersion

model $f(x;\theta,\phi) = e^{\frac{x\theta+U(\theta)+\phi V(x,\phi)}{\phi}}$ allows me to circumvent the problem of over or under dispersion. The dispersion parameter ϕ is an unknown scale parameter, earlier it was stated that in general $\text{var}(X) = \phi \frac{\partial}{\partial b} E[X]$, this indicates that statistical variance is closely related to the concept of scale. For instance the normal distribution is described in terms of two parameters, a location and a scale parameter. The location parameter corresponds to the mean μ and the scale parameter to the variance σ^2.

5.4.0 Information Theory

Likelihood is connected to Information theory, as will be seen later the likelihood function appears in the Akaike Information Criteria. Information theory may be defined as the mathematical study of methods and limits for data communication. In 1948 Claude Shannon (Shannon 1948) an American Mathematician and Electrical Engineer published a paper which may be regarded as laying the foundation of modern information theory. Information theory is a rich and fascinating area of study; statisticians owe a great deal to the work of electrical engineers and mathematicians such as Shannon. As indicated earlier some background material on information theory is useful in discussing the Akaike Information Criteria (Akaike 1974), and indeed the general problem of accessing model fit.

5.4.1 Information and Entropy

Again I consider a simple coin tossing experiment; assuming I have a fair coin, and that I toss the coin say eight times. A typical outcome would be the sequence 1 0 1 0 0 1 0 0, where 1 denotes HEADS. Now suppose I toss the coin another eight times, I observe the sequence

0 0 0 0 1 0 1 1. I repeat this operation a number of times, I build up a set of sequences such as those shown below:

1 0 1 0 0 1 0 0

0 0 0 0 1 0 1 1

1 1 1 0 0 1 1 0

0 0 0 0 1 1 0 1

1 0 1 1 0 0 0 1

0 0 0 0 1 0 1 0

0 0 1 0 1 1 0 0

0 0 0 0 1 1 0 1

Etc.

Now consider a similar experiment but this time I use a biased coin, let P(HEADS) =0.99. Then I might obtain the following set of sequences:

1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1

Etc.

The obvious difference between these two sets of sequences is that for the fair coin each sequence displays a degree of variety or variation, whereas those for the biased coin are identical. If these sequences were used to convey information then those

generated by using a fair coin would allow us to present 'richer' patterns. With the fair coin I am less certain that a HEAD will appear, but I have greater information. With the biased coin I am almost certain that a HEAD will appear, but there is a drastic reduction in the amount of information. Uncertainty and information can be measured by what is known as entropy. For a random variable X having *n* possible outcomes the Shannon information entropy $H(X)$ is given by $-\sum_{1}^{n} p(x_i)\log_b p(x_i)$.

Again using the example of a coin, figure 5.2 below is a plot of $H(X)$ versus probability of getting HEADS.

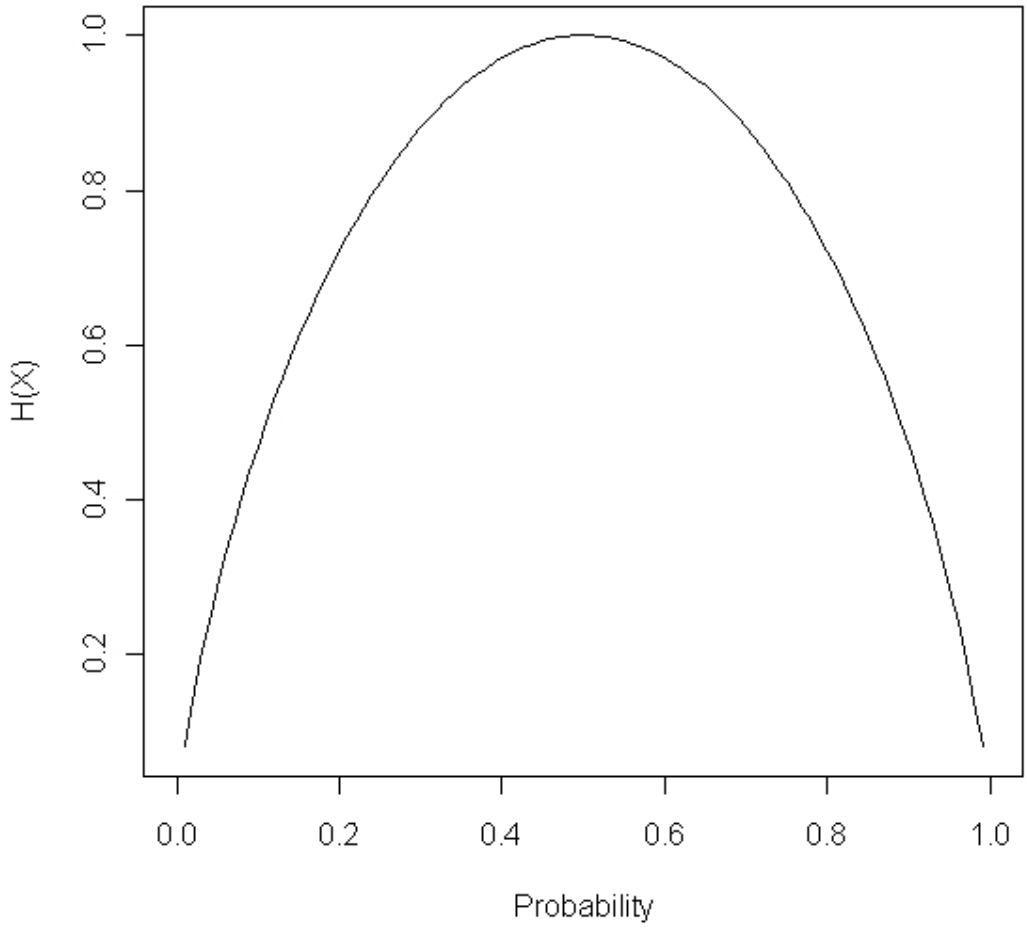


Figure 5.2 Shannon information entropy versus Probability.

It can be seen that $H(X)$ is 0 for P(HEADS)=0 and P(HEADS)=1, i.e. for sequences such as 0 0 0 0 0 0 0 0 and 1 1 1 1 1 1 1 1 the Shannon information entropy is 0, sequences such as these are easily predicted. In situations where I am certain of the outcome I find very little information is present. In situation where I am less certain of the outcome, I find greater information is present. In the coin example if P(HEADS)=0.5, then $H(X)$ is at a maximum. With a fair coin I obtain sequences such as 1 0 1 0 0 1 0 0 which is less predictable and so contains more information.

For the benefit of the interested reader additional material on entropy and statistical physics is presented in Box 1, Material on entropy and comparing probability distributions is presented in Box 2. I include this material because I believe it may provide an interesting historical background to the origins of quantities such as the AIC.

Box 1 Entropy and Statistical Physics

For the benefit of the interested reader we shall now look at the connection between Shannon information entropy and entropy as defined in statistical physics. I shall consider some standard results from thermodynamics. One the seminal papers in statistical physics was written by Ludwig Boltzmann see (Boltzmann 1872) and (Cercignani 2007). An excellent treatment of statistical physics can be found in (Blundell & Blundell 2006). For a Carnot cycle we have

$\frac{Q_e}{Q_l} = \frac{T_e}{T_l}$, where Q_e and Q_l are the heat entering and leaving the system respectively, and T_e and T_l are the temperatures of two heat reservoirs between the system, note $T_e > T_l$. Let ΔQ_{rv} be the heat entering the system at each point, then $\sum_{cycle} \frac{\Delta Q_{rv}}{T} = \frac{Q_e}{T_e} + \frac{(-Q_l)}{T_l} = 0$, we may write this in the form of an integral $\oint \frac{dQ_{rv}}{T} = 0$. Given that $\oint \frac{dQ_{rv}}{T} = 0$, then $\int_{\alpha}^{\beta} \frac{dQ_{rv}}{T}$ is independent of the path, we may express $\frac{dQ_{rv}}{T}$ as an exact differential, $dS = \frac{dQ_{rv}}{T}$, S is defined to be the entropy. The first law of thermodynamics may be stated

in the form $du = dQ + dW$. We may write $dQ = TdS$ and $dW = -pdV$, so $du = TdS - pdV$. Also using total derivatives we have

$dU = \left(\frac{\partial U}{\partial S}\right)dS + \left(\frac{\partial U}{\partial V}\right)dV$, hence $T = \frac{\partial U}{\partial S}$ and $p = -\frac{\partial U}{\partial V}$. By definition temperature *T is* given by $\frac{1}{k_B T} = \frac{d\log_e(\Omega)}{dE}$ where Ω is the number of microstates associated with a particular macrostate. By combining $\frac{1}{k_B T} = \frac{d\log_e(\Omega)}{dE}$ and $T = \frac{\partial U}{\partial S}$, we can obtain an expression for *S* as follows:

Rearranging $\frac{1}{k_B T} = \frac{d\log_e(\Omega)}{dE}$ gives $T = \frac{dE}{d\log_e(\Omega)k_B}$. So we have $\frac{\partial U}{\partial S} = \frac{dE}{d\log_e(\Omega)k_B}$, hence $\frac{\partial S}{\partial U} = \frac{d\log_e(\Omega)k_B}{dE}$.

Integrating we obtain $S = k_B \log_e(\Omega)$, this is the Boltzman expression for entropy.

Let a system have an number of equally likely states N_{ob} , then the entropy *S* is $k_B \log_e(N_{ob})$. However it may be that each of the N_{ob} states comprises of a number of microstates, which may be extremely difficult to observe or measure, the total entropy $S_t = S + S_m$, where S_m is the entropy connected with the microstates. Let a system have *N* equally likely microstates, if theses microstates are arranged into groups (macrostates) with N_i microstates contained within the ith macrostate, then $\sum_i N_i = N$. The probability P_i that the system occupies the ith macrostates is given by $P_i = \frac{N_i}{N}$. Now $S = S_t - S_m$, S is the measurable entropy. We have $S_t = k_B \log_e(N)$, and the entropy of the microstates within the ith macrostate is $S_i = k_B \log_e(N_i)$. We cannot measure S_m the entropy connected with being in any different microstate. However we can access S_m , through the relationship $S_m = E(S_i)$, (note here *E* denotes the expected value), so $S_m = E(S_i) = \sum_i k_B P_i \log_e(N_i)$. From $S = S_t - S_m$ we see that $S = k_B \log_e(N) - \sum_i k_B P_i \log_e(N_i)$, this expression may written as

$k_B \sum_i P_i(\log_e(N) - \log_e(N_i))$ or $k_B \sum_i P_i \log_e\left(\frac{N}{N_i}\right) = -k_B \sum_i P_i \log_e(P_i)$.

$S = -k_B \sum_i P_i \log_e(P_i)$ (The Shannon entropy), we see here the similarity in functional form of the Shannon entropy and the Shannon information entropy $H(X) = -\sum_1^n p(x_i) \log_b p(x_i)$.

Box 2 Entropy and the Comparison of Probability Distributions

We notice that in the expression for both the Shannon entropy and the Shannon information entropy we are dealing with one probability distribution, let us assume this is the true distribution and denote the p.d.f by $p(x)$. Let us consider some other distribution with p.d.f $q(x)$. An important question would be how different is $q(x)$ from $p(x)$? The expression

$-k_B \sum_i p_i(x) \log_e \left(\frac{q_i(x)}{p_i(x)} \right) = -k_B \sum_i p_i(x) \log_e q_i(x) - p_i(x) \log_e p_i(x)$ can be written as

$-k_B (E[\log_e q_i(x)] - E[\log_e p_i(x)])$.

The quantity $(E[\log_e q_i(x)] - E[\log_e q_i(x)])$ provides us with a measure of the 'difference' or distance between $q(x)$ and $p(x)$. If $q(x)$ is close to the true distribution $p(x)$ then $(E[\log_e q_i(x)] - E[\log_e p_i(x)])$ will be small.

We can describe

$-k_B \sum_i p_i(x) \log_e \left(\frac{q_i(x)}{p_i(x)} \right) = -k_B \sum_i p_i(x) \log_e q_i(x) - p_i(x) \log_e p_i(x)$ as the generalised Boltzmann entropy , denoted GB (Chakrabarti & Chakrabarty 2006) . The quantity $(E[\log_e p_i(x)] - E[\log_e q_i(x)])$ is of particular importance in statistics as it relates closely to the Kullback Leibler distance, denoted KL, see (Kullback & Leibler 1951), (Bozdogan 1987) and (Nariaki 1978). For a discrete random variable we have KL given by $\sum_i p_i(x) \log_e \left(\frac{p_i(x)}{q_i(x)} \right)$, we see that except for the constant k_B, $KL = -GB$. For a continuous random variable we have KL given

by $\int_{\alpha}^{\beta} p(x)\log_e\left(\frac{p(x)}{q(x)}\right)dx = \int_{\alpha}^{\beta} p(x)\log_e\ p(x)dx - \int_{\alpha}^{\beta} p(x)\log_e\ q(x)dx$. The first term in the previous expression is the Shannon entropy which is constant, the second term $-\int_{\alpha}^{\beta} p(x)\log_e\ q(x)dx$ is known as the cross entropy. The cross entropy gives us a measure of the distance between $p(x)$ and $q(x)$. Viewing KL as a measure of the distance between the true distribution $p(x)$ and $q(x)$, we need to minimise the cross entropy.

The following example may help us to see what the KL is about. Let the true distribution $p(x)$ of X be the standard normal distribution. So $p(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x}{2}}$.

Let $q(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, then using $\int_{-\infty}^{\infty} p(x)\log_e\left(\frac{p(x)}{q(x)}\right)dx$ we have

$$KL = \log_e(\sqrt{2\pi}\sigma) - \log_e(\sqrt{2\pi}) + \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\frac{(x-\mu)^2}{2\sigma^2}dx - \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\frac{x^2}{2}dx.$$

We may write, $KL = \log_e(\sqrt{2\pi}\sigma) - \log_e(\sqrt{2\pi}) + \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\left(\frac{(x-\mu)^2}{2\sigma^2} - \frac{x^2}{2}\right)dx.$

So

$$KL = \log_e(\sqrt{2\pi}\sigma) - \log_e(\sqrt{2\pi}) + \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\left(\frac{x^2 - 2\mu x + \mu^2}{2\sigma^2} - \frac{x^2}{2}\right)dx.$$

Also the integral in the above expression can be written

$$\frac{1}{2\sigma^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}x^2dx - \frac{\mu}{\sigma^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}xdx + \frac{\mu^2}{2\sigma^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x}{2}}dx - \frac{1}{2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}x^2dx$$

This may be written as

$$\left(\frac{1}{2\sigma^2} - \frac{1}{2}\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}x^2dx - \frac{\mu}{\sigma^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}xdx + \frac{\mu^2}{2\sigma^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx.$$

Using the fact that for a continuous random variable $\sigma^2 = \int_{\alpha}^{\beta} p(x)x^2 - \mu^2$, we may

write $\int_{\alpha}^{\beta} p(x)x^2 dx = \sigma^2 + \mu^2$, we know that X follows the standard normal distribution, so

$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} x^2 dx = 1$. Also $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} x dx = 0$ and by definition

$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1$, so $\frac{\mu^2}{2\sigma^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{\mu^2}{2\sigma^2}$.

Thus

$$\left(\frac{1}{2\sigma^2} - \frac{1}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} x^2 dx - \frac{\mu}{\sigma^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} x dx + \frac{\mu^2}{2\sigma^2} = \frac{1}{2\sigma^2} - \frac{1}{2} + \frac{\mu^2}{2\sigma^2}.$$

So $KL = \log_e(\sqrt{2\pi}\sigma) - \log_e(\sqrt{2\pi}) + \frac{1}{2\sigma^2} + \frac{\mu^2}{2\sigma^2} - \frac{1}{2} = \log_e(\sigma) - \frac{1}{2}\left(\frac{1+\mu^2}{\sigma^2} - 1\right)$

As a rule we of course do not know what the true distribution is. If the true distribution was for example the normal distribution with mean μ_t and variance $\sigma^2{}_t$, and we have $q(x)$ as above, then

$KL = \frac{1}{2}\log_e\left(\frac{\sigma^2{}_t}{\sigma^2}\right) + \frac{1}{2}\left(\frac{\sigma^2}{\sigma^2{}_t} - 1 + \frac{(\mu_t - \mu)}{\sigma^2{}_t}\right)$. We see here that KL is expressed in terms of μ_t and $\sigma^2{}_t$.

We can make the following remarks about KL:

$KL \geq 0$ and if $p(x) \neq q(x)$ then $KL(p;q) \neq KL(q;p)$ i.e. KL is not symmetric.

The first result can be obtained as follows:

Let $u = \frac{q(x)}{p(x)}$.

Jensen's inequality states

$E[-\log_e(u)] \geq -\log_e E[u]$, where *u* is a convex function.

So we have

$$E[-\log_e\left(\frac{q(x)}{p(x)}\right)] \geq -\log_e E[\frac{q(x)}{p(x)}]$$

Therefore

$$\int_{\alpha}^{\beta} -\log_e\left(\frac{q(x)}{p(x)}\right)p(x)dx \geq -\log_e\left(\int_{\alpha}^{\beta}\left(\frac{q(x)}{p(x)}\right)p(x)dx\right)$$

But $\int_{\alpha}^{\beta}\left(\frac{q(x)}{p(x)}\right)p(x)dx = 1$.

So $\int_{\alpha}^{\beta} -\log_e\left(\frac{q(x)}{p(x)}\right)p(x)dx \geq 0$

Hence

$$\int_{\alpha}^{\beta} \log_e\left(\frac{p(x)}{q(x)}\right)p(x)dx \geq 0$$

Now $KL = \int_{\alpha}^{\beta} \log_e\left(\frac{p(x)}{q(x)}\right)p(x)dx$, so

$KL \geq 0$.

We can show $KL(p;q) \neq KL(q;p)$ as follows:

$KL(p;q) = \int_{\alpha}^{\beta} \log_e\left(\frac{p(x)}{q(x)}\right)p(x)dx$ and $KL(q;p) = \int_{\alpha}^{\beta} \log_e\left(\frac{q(x)}{p(x)}\right)q(x)dx$, so

$$KL(p;q) = \int_{\alpha}^{\beta} \log_e\, p(x)\,p(x)dx - \int_{\alpha}^{\beta} \log_e\, q(x)\,p(x)dx$$

$$KL(q;p) = \int_{\alpha}^{\beta} \log_e\, q(x)q(x)dx - \int_{\alpha}^{\beta} \log_e\, p(x)q(x)dx$$

$KL(p;q) = KL(q;p)$ if and only if $\int_{\alpha}^{\beta} \log_e\, p(x)\,p(x)dx = \int_{\alpha}^{\beta} \log_e\, q(x)q(x)dx$.

There is a significant drawback to KL, we cannot observe it. KL relies on us knowing the true distribution. As stated earlier KL is expressed in terms of the parameters of the true distribution which are unknown. We need then to consider how we might obtain an estimate of KL from the data.

5.4.2 Estimating theKullback Leibler distance by the AIC

When fitting models to a dataset I desire the model which maximises the likelihood (or the log likelihood). That is, I want to maximise $\log_e L(\theta) = \sum_i^n \log_e(f(x_i))$. I could also look at maximising

$$\frac{\log_e L(\theta)}{n} = \frac{\sum_i^n \log_e(f(x_i))}{n}.$$

For a large enough dataset

$$\frac{\sum_i^n \log_e(f(x_i))}{n} = E[\log_e(f(X)].$$

So in maximising the likelihood I maximise $E[\log_e(f(X)]$ (note *f()* is the model that I am trying to fit). In my discussion of the Kullback Leibler (KL) distance I noted that I aim to minimise the cross entropy (and so minimise KL) i.e. minimise

$$-\int_\alpha^\beta p(x)\log_e q(x)dx = -E[\log_e q(x)] \text{ or maximise} \int_\alpha^\beta p(x)\log_e q(x)dx = E[\log_e q(x)].$$

KL can be estimated by $\int_\alpha^\beta p(x)\log_e p(x)dx - \frac{\sum_i^n \log_e(f(x_i))}{n}$. I said earlier that the Shannon entropy is constant; this means that KL can be estimated by

$$-\frac{\sum_i^n \log_e(f(x_i))}{n}$$

I can make the following important statements.

1 Maximising the likelihood is equivalent to minimising KL, i.e. maximising $E[\log_e(f(X)]$ is equivalent to minimising KL.

2 In fitting models to a dataset I seek the model that maximises $E[\log_e(f(X)]$.

In practical situations I would maximise $E[\log_e f(X;\hat{\theta})]$ where $\hat{\theta}$ is a MLE of the parameter θ, (note $\hat{\theta}$ and θ could of course be vectors). It is tempting to think that I can use $\sum_i^n \log_e(f(x_i);\hat{\theta})$, the maximised log likelihood, to estimate $E[\log_e f(X;\hat{\theta})]$, but this quantity is biased. For instance if I have nested models; the model with the largest number of parameters will always give the largest value for $\sum_i^n \log_e(f(x_i);\hat{\theta})$. Similarly this problem with bias means that estimating KL by $-\frac{\sum_i^n \log_e(f(x_i))}{n}$ leads to a distorted estimate of KL.

There are several ways to tackle the problem of obtaining an unbiased estimate of $E[\log_e f(X;\hat{\theta})]$, one example is The Jackknife method (Miller 1974). For example the jackknife method can be used to obtain an estimate of standard error in the following way. Let $\bar{x}_{j\neq i}$ be the sample mean of based on the sample with the *ith* observation deleted. Let $\bar{x}_{()}$ be the average of $\bar{x}_{j\neq i}$ .Then $\left[\sum_{j=1}^n \frac{n-1}{n}(\bar{x}_{j\neq i}-\bar{x}_{()})^2\right]^{\frac{1}{2}}$ is the jackknife estimate of the standard error.

Another method is the Akaike Information Criteria (AIC). I shall alter slightly the notation and use $E[\log_e f_k(X;\hat{\theta}_k)]$, this is to remind me that I am considering the k[th] model from a number of possible models. The AIC is given by $AIC(k) = -2\sum_i^n \log_e f_k(x_i;\hat{\theta}_k) + 2p$, where p is the number of parameters in the model. The AIC is an unbiased estimate of $-2nE[\log_e f_k(X;\hat{\theta}_k)]$, therefore I wish to

find the model that minimises the AIC. I shall now look at how the AIC is derived; from the derivation described in chapter 13 of Pawitan's book (Pawitan 2001). Given the model $f_k(x,\theta_k)$, I have $\log_e L(\theta_k) = \sum_i^n \log_e f_k(x_i;\theta_k)$. Let the solution of $E[\log_e f(X;\theta_k)] = 0$ be θ_{kS} , estimate θ_{kS} by $\hat{\theta}_k$, (θ_{kS} and $\hat{\theta}_k$ are vectors). Now define $J_k \equiv E\left(\frac{\partial \log_e f_k(X,\theta_k)}{\partial \theta_k}\right)\left(\frac{\partial \log_e f_k(X,\theta_k)}{\partial \theta_k'}\right)$ and $I_k = -E\frac{\partial^2 \log_e f(X,\theta_k)}{\partial \theta_k \partial \theta_k'}$. I will need to make use of the result $E[n(\hat{\theta}_k - \theta_{kS})' I_k (\hat{\theta}_k - \theta_{kS})] \approx tr(J_k I_k^{-1})$. (Note A' denotes the transpose of a matrix, and *tr* is the trace of a matrix, i.e. the sum of the elements in the main diagonal). The Taylor series for $\log_e L(\theta_k) = \sum_i^n \log_e f_k(x_i;\theta_k)$ about $\hat{\theta}_k$ is

$$\log_e L(\theta_k) = \log_e L(\hat{\theta}_k) + \frac{\partial \log_e L(\hat{\theta}_k)}{\partial \theta_k}(\theta_k - \hat{\theta}_k) + \frac{1}{2}(\theta_k - \hat{\theta}_k)' \frac{\partial^2 \log_e L(\theta^*_k)}{\partial \theta_k \partial \theta_k'}(\theta_k - \hat{\theta}_k) + \ldots$$

, where $\left|\theta_k^* - \theta_k\right| \le \left|\theta_k - \hat{\theta}_k\right|$.

For large samples $\hat{\theta}_k \to_{prob} \theta_k$, so $E\left(\frac{\partial \log_e L(\hat{\theta}_k)}{\partial \theta_k}\right) = 0$. I now have the approximation $\log_e L(\theta_k) = \log_e L(\hat{\theta}_k) + \frac{1}{2}(\theta_k - \hat{\theta}_k)' \frac{\partial^2 \log_e L(\theta^*_k)}{\partial \theta_k \partial \theta_k'}(\theta_k - \hat{\theta}_k)$.

Again appealing to large samples results I have

$\frac{1}{n}\frac{\partial^2 \log_e L(\theta^*_k)}{\partial \theta_k \partial \theta_k'} \to_{prob} E\left(\frac{\partial^2 \log_e f(X;\theta_k)}{\partial \theta_k \partial \theta_k'}\right) = -I_k$, which gives

$$\frac{\partial^2 \log_e L(\theta^*_k)}{\partial \theta_k \partial \theta_k'} \to_{prob} -nI_k .$$

I may now write $\log_e L(\theta_k) = \log_e L(\hat{\theta}_k) + \frac{1}{2} n(\theta_k - \hat{\theta}_k)' I_k (\theta_k - \hat{\theta}_k)$, so with $\theta_k = \theta_{ks}$

and using $E[n(\hat{\theta}_k - \theta_{kS})' I_k (\hat{\theta}_k - \theta_{kS})] \approx tr(J_k I_k^{-1})$,I have

$\log_e L(\theta_{ks}) \approx \log_e L(\hat{\theta}_k) - \frac{1}{2} tr(J_k I^{-1}{}_k)$. Therefore

$E[\log_e L(\theta_{ks})] \approx E[\log_e L(\hat{\theta}_k)] - \frac{1}{2} tr(J_k I^{-1}{}_k)$. Using the fact that

$\frac{1}{n} \log_e L(\theta_{ks}) \rightarrow_{prob} E[\log_e f(X;\theta_{ks})]$, I arrive at the approximation

$nE[\log_e f(X;\theta_{ks})] \approx E[\log_e L(\hat{\theta}_k)] - \frac{1}{2} tr(J_k I^{-1}{}_k)$. The Taylor series for

$E[\log_e f(X;\theta_k)]$ about θ_{ks} is

$$E[\log_e f(X;\theta_{ks})] + \frac{\partial E[\log_e f(X;\theta_{ks})]}{\partial \theta_k} (\theta_k - \theta_{ks}) + \frac{1}{2} (\theta_k - \theta_{ks})' \frac{\partial^2 E[\log_e f(X;\theta^*{}_k)]}{\partial \theta_k \partial \theta_k'} (\theta_k - \theta_{ks}) + \ldots$$

I then obtain the approximation

$$E[\log_e f(X;\theta_k)] \approx E[\log_e f(X;\theta_{ks})] - \frac{1}{2} (\theta_k - \theta_{ks})' I_k (\theta_k - \theta_{ks}).$$

Setting $\theta_k = \hat{\theta}_k$ I get

$$E[\log_e f_k(X;\hat{\theta}_k)] = E[E[\log_e f(X;\hat{\theta}_k)]] \approx E[\log_e f(X;\theta_{ks})] - \frac{1}{2n} tr(J_k I^{-1}{}_k).$$

On combining $nE[\log_e f(X;\theta_{ks})] \approx E[\log_e L(\hat{\theta}_k)] - \frac{1}{2} tr(J_k I^{-1}{}_k)$ and

$E[\log_e f_k(X;\hat{\theta}_k)] = E[E[\log_e f(X;\hat{\theta}_k)]] \approx E[\log_e f(X;\theta_{ks})] - \frac{1}{2n} tr(J_k I^{-1}{}_k)$, I get

$nE[\log_e f(X;\hat{\theta}_k)] \approx E[\log_e L(\hat{\theta}_k)] - tr(J_k I^{-1}{}_k)$. From the last result I can say that

$\log_e L(\hat{\theta}_k) - tr(J_k I^{-1}{}_k)$ is an unbiased estimator of $nE[\log_e f(X;\hat{\theta}_k)]$. The AIC is

based on the assumption that $J_k = I_k$, this means that $tr(J_k I^{-1}{}_k)$ is approximately

equal to the number of parameters in the model, i.e. $tr(J_k I^{-1}{}_k) \approx p$. So I have

$\log_e L(\hat{\theta}_k) - p$ is an unbiased estimator of $nE[\log_e f(X;\hat{\theta}_k)]$, hence $AIC(k) = -2\log_e L(\hat{\theta}_k) + 2p$ is an unbiased estimator of $-2nE[\log_e f(X;\hat{\theta}_k)]$. The AIC is an estimator of $2E[KL]$ (Bozdogan 1987). The first term in the AIC formula gives a measure of how bad a fit a particular model is to the data. The AIC penalises model complexity through p, I said earlier that when fitting models I look for the model that gives the smallest value for the AIC, as a model becomes more complex p increases and so the AIC increases. It should be noted that single values of the AIC are not of use to me in fitting a model, I must look at changes in the AIC. For example I might examine the changes in the AIC when a new term is introduced into a model. A simple example might be to consider the models $\beta_0 + \beta_1 X_1$ and $\beta_0 + \beta_1 X_1 + \beta_2 X_2$, let AIC_1 be the AIC for the first model, and AIC_2 be the AIC for the second model. Then I look at $AIC_1 - AIC_2$. In the past there has been some interesting discussion concerning the term *2p* in the AIC formula. Questions have been raised regarding the adequacy of penalisation as implemented in the AIC (Bozdogan 1987). Put simply is 2 a big enough multiplier? The AIC hinges on the approximation $tr(J_k I^{-1}{}_k) \approx p$, if this approximation is not correct, then the AIC will not give an unbiased estimate of $-2nE[\log_e f(X;\hat{\theta}_k)]$. Concern has been expressed over the question of the consistency of the estimate of model order (*k*) obtained through minimising the AIC. As the sample size increases the order of the best model obtained by using the AIC will increase, however it may not be close to the order of the 'true' model. As far as the AIC is concerned the true model could be of infinite order. It is important to note that the AIC can be used to select the best fitting model but not as a means of estimating the true order of the model. This last issue is of interest in that it relates to possible over-fitting or under-fitting. Bozdogan (Bozdogan 1987) argues that

'consistency is an asymptotic property and any real problem has a finite sample size n'. Bozdogan also makes an extremely important remark to the effect that consistency supposes that there is a 'true' model order. I would argue that the AIC is a very good model selection tool; it is attractive due to its relationship to the fundamental measure KL.

5.5.0 Extending the AIC

The AIC in the form that I have considered is for want of a better description the 'classical' form. With some effort I can see the connection between the AIC and KL. The 'classical' form of the AIC has some limitations. In this present work one of my main concerns is the problem of over-fitting. An important question is, if I use the AIC as a model selection tool am I liable to over-fit models? In certain circumstance the answer to this question is yes. Using the formula $AIC = -2\log_e L(\hat{\theta}) + 2p$, what happens if p large is compared to the sample size? In this case I will find that models selected using this form of the AIC are prone to over-fitting. To overcome the problem of over-fitting when the sample size n is small compared to p, I have to consider a corrected version of the AIC. This corrected AIC is denoted AIC_c, and is given by the formula

$AIC_c = -2\log_e L(\hat{\theta}) + 2p + \frac{2p(p+1)}{n-p-1}$, see (Nariaki 1978) and (Hurvich & Tsai 1995). It is seen that as $n \to \infty$, $AIC_c \to AIC$, it is suggested that AIC_c be used as opposed to AIC in situations where $\frac{n}{p} < 40$,see (Burnham & Anderson 2004). Can I use the AIC in situations where I might want to fit a mixed model, a model with time dependent covariates or a frailty model? I shall now consider these three cases.

5.5.1 Mixed Models

I make reference to mixed models in Chapter 8, briefly the model $Y_{jk} = \beta_0 + \zeta_j + (\alpha_j + \beta_1)X_k$ is known as a mixed model. The model contains fixed and random effects, ς_j is a random intercept and α_j is a random slope. Random effects are handled in the same way as the fixed effects, so that I have $p + p_j$ parameters in the model (p fixed and p_j random). Let $\hat{\theta}$ be a vector, this gives $AIC = -2\log_e L(\hat{\theta}) + 2(p + p_j)$ and

$AIC_c = -2\log_e L(\hat{\theta}) + 2(p + p_j) + \frac{2(p + p_j)(p + p_j + 1)}{n - (p + p_j) - 1}$. Earlier I mentioned the problem of the consistency of the AIC, a consistent form of the AIC is $CAIC = -2\log_e L(\hat{\theta}) + p((\log_e n) + 1)$, see (Bozdogan 1987). Again with $p + p_j$ total number of parameters, I have $CAIC = -2\log_e L(\hat{\theta}) + (p + p_j)((\log_e n) + 1)$. In terms of actually fitting a mixed model, the AIC, AIC_c and CAIC are implemented in a number of statistical software packages. For example in the SAS procedures MIXED and GLMMIX various forms of the AIC and related quantities are implemented:

$AIC = -2l + 2d$

$AICC = -2l + \frac{2dn^*}{(n^* - d - 1)}$

$HQIC = -2l + 2d\log_e(\log_e n)$

$BIC = -2l + d\log_e n$

$CAIC = -2l + d(\log_e n + 1)$

where

$l = -2\log_e L(\hat{\theta})$

In the above formulae $d = p + p_j$. The MIXED procedure uses restricted maximum likelihood (sometimes known as residual maximum likelihood) in fitting a model. The BIC (Bayesian Information Criterion) was developed by Schwarz in 1978 (Schwarz 1978), Burnham (Burnham & Anderson 2004) contains a most interesting discussion on the BIC, as Burham points out the BIC unlike the AIC is not related to information theory. In deriving the BIC unlike the AIC, it is not assumed that the model used in the derivation is the 'true' model.

Buckland et al. (Buckland *et al.* 1997) state that the BIC is consistent in terms of the dimensions of the best models selected; the BIC assumes that the 'true' model is of small dimension. For small samples the BIC is prone to select models that are under-fitted. Taking this into consideration, in regard to over-fitting would we do better to use the BIC rather than the AIC? This question is not at all straightforward, whether to use the BIC or the AIC depends upon the dimensions of the 'true' model. The BIC might have advantages over the AIC if the underlying 'true' model is of low dimension. Buckland et al. (Buckland *et al.* 1997) make some very astute remarks in regard to the question of whether it is better to use the AIC or the BIC. I would say that there is no grave disadvantage in using the AIC, however the issues raised in Buckland et al. are thought provoking and I would find further investigation of this question fascinating.

When using the AIC to gauge the fit of a mixed model the researcher should exercise some degree of caution. Vaida and Blanchard (Vaida & Blanchard 2005) consider clustered data, they show that the AIC in its classical form leads to rather strange results when applied for example to repeated measures. They develop a conditional AIC. Vaida and Blanchard consider an example using repeated measures, see (Vaida

& Blanchard 2005), the data consists of six measurements taken on ten patients, they use the nlme package in GNU R to produce a mixed model and a linear regression model. On comparing the AIC they find that the linear regression model is favoured over the mixed model, this they point out is strange given that the linear regression model has 21 parameters and the mixed model 6. Essentially their argument is that the AIC in its classical form when applied to mixed models leads to misleading results because the penalty term is not appropriate for the mixed model situation (penalty term is too large). Consequently they develop a form of the AIC with an adjusted penalty term appropriate for use in developing a mixed model. It should be noted that the nlme package has been superseded by lmer4. SAS's PROC MIXED implements an appropriate form of the AIC (Fernandez 2007). It is advisable that researchers try to establish which form of the AIC is implemented in the particular software package they happen to be using.

5.5.2 Time Dependent Covariates

When analysing survival data with say the Cox Proportional Hazards model, it is often assumed that covariates do not change with time. A covariate is taken as remaining constant up to the event of interest. The Cox Proportional Hazards model can be extended by considering covariates that change with time, time dependent covariates. The SAS procedure PHREG allows one to fit a model with time dependent covariates. PHREG reports the AIC and BIC (note the BIC is reported as SBC, Schwarz Bayesian Criterion).

5.5.3 Frailty Models

Do Ha et al. (Do Ha *et al.* 2007) develop an AIC for a set of frailty models, (the models need not be nested). The AIC proposed by Do Ha *et al* is based on conditional likelihood and an extended restricted likelihood. They define two AICs as follows:

$AIC(D^*) = D^* + 2p_D^*$

$AIC(T_d^*) = T_d^* + 2p_T$

$AIC(D^*)$ deals with fixed and random effects and frailty parameters, whilst $AIC(T_d^*)$ deals only with the frailty parameters (dispersion parameters in the frailty distribution). In (Do Ha *et al.* 2007) the authors state that for a Cox Proportional Hazards model $AIC(D^*)$ is the AIC as used in the SAS procedure PHREG. For a linear mixed model $AIC(T_d^*)$ is the AIC as used in the SAS procedure MIXED. The authors also suggest that in regard to frailty terms $AIC(T_d^*)$ may be a better selection criterion than $AIC(D^*)$. The work of Do Ha *et al.* is certainly very interesting; as far as I am aware the AICs developed in (Do Ha *et al.* 2007) have not been implemented as software. Further investigation of the results presented by Do Ha *et al.* would be well worth pursuing. In the next chapter will I shall further consider over-fitting and model optimism, examine Harrell's C and discuss the issue of validation techniques.

CHAPTER 6 OVER-FITTING, OPTIMISM AND VALIDATION

- When trying to develop a prognostic model including a large number of potential predictor variables may lead to over-fitting
- Over-fitted models are biased in regard to predictive power
- Over-fitted models are poor prognostic tools
- Prognostic models should be validated

6.0.0 Introduction

I have made reference to over-fitting numerous times in the preceding chapters. I could define over-fitting to be the tendency in certain statistical modelling procedures to produce models that include substantial noise, that is I end up with a model that does not just describe the general patterns in a data set, but includes a deal of local fine-grained detail. Over-fitting leads to models that include variables that are significant in the sense that they model local detail, they may not be significant as general overall predictors. A model that has been over-fitted is biased in terms of how optimistic predictions based on this model .If over-fitting is present then on applying the model to a new but similar data set I would see a change (deterioration) in the predictive power of the model when used to predict on the new data set. This difference in predictive power can be described in terms of optimism. I can gauge the predictive power of a model by measuring the agreement between the observed and predicted values of the dependent variable. One way of measuring such agreement is by Somer's D (Somers 1962). Harrell (Harrell *et al.* 1996) defines optimism in terms

of the difference in two values of Somer's D. It may be worthwhile looking at Somer's D and some related measures.

6.1.0 Somer's D

The population value of Somer's D is defined as follows $D_{XY} = \frac{\tau_{XY}}{\tau_{XX}}$,

where $\tau_{XY} = E[\text{sgn}(X_i - X_j)\text{sgn}(Y_i - Y_j)]$, for all i, j, τ_{XY} is Kendall's τ_a (Kendall 1938).

The sgn function is defined as follows:

$\text{sgn}(x) = -1, x < 0$, $\text{sgn}(x) = 0, x = 0$, $\text{sgn}(x) = 1, x > 0$.

X and Y are sampled jointly from a bivariate distribution.

Kendall's τ_a gives a measure of concordance, the $X's$ and $Y's$ are said to be concordant if the bigger of the $X's$ is associated with the bigger of the $Y's$. Somer's D is the regression coefficient of $\text{sgn}(X_i - X_j)$ with respect to $\text{sgn}(Y_i - Y_j)$. Both Kendall's τ_a and Somer's D can be applied to survival data, X or Y or both could be censored. If I have indicator variables U and V, where values of 1 indicate that the event of interest has occurred and values of 0 indicate censoring, then Somer's D for survival data can be defined as $D_{Y,V,X,U} = \frac{\tau_{X,U,Y,V}}{\tau_{X,U,X,U}}$.

6.1.2 Harrell's C

Harrell (Harrell *et al.* 1996) has defined the quantity $D_{X,1,Y,S} = 2C - 1$, where it is assumed that X is a continuous variable. C is known as Harrell's C. I can interpret Harrell's C as measuring how well X predicts survival. Harrell's C is defined as

"the proportion of all usable patient pairs in which the predictions and outcomes are concordant" (Harrell *et al.* 1996). For a binary dependent variable C is "the proportion of all pairs of patients, one with and one without the disease, in which the patient having the disease had a higher predicted probability of disease." (Harrell *et al.* 1996). Harrell's C takes on values from 0.5 to 1.0, 0.5 indicating poor predictive power (poor level of agreement between predicted and observed $Y's$) and 1.0 indicating very good predictive power (high level of agreement between predicted and observed $Y's$). Somer's D can take on values from -1 to 1.

Assuming then that I have arrived at my final model, I could use Somer's D or Harrell's C to obtain some measure of the predictive power of the model. However, I will not address the problem of possible over-fitting by examining values of Somer's D or Harrell's C for the original data set alone. Ideally to assess over-fitting I need to fit the model to a number of different but similar data sets and examine predictive power of the model over these data sets. This leads to the idea of validation methods.

6.2.0 Validation Methods

The formal procedure for determining the predictive power or accuracy of the final model is known as model validation.

I have said that in order to assess over-fitting I need to fit the 'final' model to a number of data sets that are similar but different to the data set used to develop the model. How do I obtain these data sets? I could reserve some of our original data and use it to test the model, or I could try to 'build' some data. The first approach is known as data splitting, see (Picard & Berk 1990). The second approach could involve the use of the bootstrap (Efron & Gong 1983). I shall now consider these two approaches in a little more detail.

6.2.1 Data Splitting

The idea of splitting up the original data set into a portion on which to develop the model (the training sample) and a portion for validation seems quite reasonable. So how is the data split up? This question is not trivial and Picard and Berk (Picard & Berk 1990) draw attention to the problem which may result if the data is split in an arbitrary way. I may end up with not enough data to develop the model, or conversely, if I reserve a large portion of the data for development, I may not have sufficient data for validation. A formal criterion for partitioning the data would be desirable, but it is often the case that the mathematical expressions for these criteria are intractable.

Picard and Berk (Picard & Berk 1990) suggest that between $\frac{1}{4}$ to $\frac{1}{2}$ of the data should be reserved for validation.

If I have a large data set I could consider repeated data splitting, this is called cross-validation (Stone 1974). With cross-validation I have multiple models (a model per split), if I have split the original data set k times, I have k training samples and k validation samples. I develop and validate the k models and then 'average' the results, i.e. I could obtain averages for regression coefficients and Somer's D. Data splitting and cross-validation tend to produce highly variable estimates. In data splitting I might see notable variation in, say, the estimate of regression coefficients dependent on how I split the original data. In cross-validation the same problem arises due to the multiple training and validation samples used. In both data splitting and cross-validation the accuracy of the estimates is highly variable. A way of overcoming this problem is to use the bootstrap.

6.2.2 The Bootstrap

The bootstrap (Efron 1979) was devised by Bradley Efron as an extension of the jackknife (Miller 1974), (Efron & Stein 1981), (Efron 2000). Efron described the bootstrap as "a muscularized big brother to the Quenoille-Tukey jackknife" (Efron 2000). The bootstrap method is described as follows.

Suppose I have a data set $\{x_1, x_2, ...x_{10}\}$.

I can form the bootstrap sample by drawing at random and with replacement from original data set. The bootstrap sample is usually written in the form $X_1^*, X_2^*, ...X_n^*$, where n is the size of the original data set. A typical bootstrap sample from $\{x_1, x_2, ...x_{10}\}$ might be $\{x_3, x_3, x_7, x_5, x_2, x_{10}, x_1, x_8, x_2, x_4\}$. I may for example want to obtain an estimate of the true standard error for some quantity or statistic, let this quantity be $\hat{\theta}$.

I use the bootstrap sample to obtain $\hat{\theta}^*$ the bootstrap replication of $\hat{\theta}$.

$\hat{\theta}^*$ is often written as $\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, ...X_n^*)$.

now from another bootstrap sample and obtain the bootstrap replicate using this sample. I repeat this process B times where B is a large number. I now have B bootstrap replicates $\hat{\theta}^*$. As $B \to \infty$ the quantity $\left[\frac{\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}^{*\cdot})^2}{(B-1)}\right]^{\frac{1}{2}}$ tends toward $\hat{\sigma}_B$ the bootstrap estimate of the standard error of $\hat{\theta}$, where $\hat{\theta}^{*\cdot} \equiv \frac{\sum \hat{\theta}^{*b}}{B}$. One extremely important feature of the bootstrap is that I do not have to know what distribution the original data comes from. The true standard error of $\hat{\theta}$, $\sigma(F)$, depends upon knowing

what distribution the original data comes from, the bootstrap allows me to estimate $\sigma(F)$ by means of $\hat{\sigma}_B$. The bootstrap estimate of $\sigma(F)$, $\hat{\sigma}_B$ depends upon the empirical distribution $\hat{F}$, so I can write $\hat{\sigma}_B = \sigma(\hat{F})$. The empirical distribution $\hat{F}$ assigns equal probability (probability mass $\frac{1}{n}$) to each x in the original data set. The bootstrap can be applied to quite complicated statistics with ease. I described B as a large number, values of B do not have to be huge, values of 200 or 300 can produce good estimates. An excellent discussion of the bootstrap and jackknife can be found in (Efron & Gong 1983). The bootstrap is an internal validation method, as are data splitting and cross-validation; a portion of the original data set is used to validate the final model. A more rigorous validation procedure would involve entirely new data sets, this may not be a practical approach, it might be difficult to obtain new data for a variety of reasons, for instance financial constraints, data collection may take a long time.

Harrell (Harrell *et al.* 1996) recommends the bootstrap as a method of internal validation; the estimates of the predictive accuracy of a model produced by the bootstrap are virtually unbiased. One major benefit of the bootstrap is that unlike data splitting and cross-validation all of the data is used to develop the model. I now consider a validation procedure as described by Harrell in (Harrell *et al.* 1996).

6.3.0 Harrell's validation procedure

In (Harrell *et al.* 1996) Harrell lists the following steps needed in order to assess the internal validity of a model. These steps are given in Box 1 exactly as they appear in (Harrell *et al.* 1996).

Box 1 Harrell's Validation Steps

1. **Develop the model using all n subjects and whatever stepwise testing is deemed necessary. Let D_{app} denote the apparent D from this model, i.e., the rank correlation computed on the same sample used to derive the fit.**
2. **Generate a sample of size n with replacement from the original sample (for both predictors and the response).**
3. **Fit the full or possibly stepwise model, using the same stopping rule as was used to derive D_{app}.**
4. **Compute the apparent D for this model on the bootstrap sample with replacement. Call it D_{boot}.**
5. **'Freeze' this reduced model, and evaluate its performance on the original dataset. Let D_{orig} denote the D.**
6. **The optimism in the fit from the bootstrap sample is $D_{boot} - D_{orig}$.**
7. **Repeat steps 2 to 6 100-200 times.**
8. **Average the optimism estimates to arrive at O.**
9. **The bootstrap corrected performance of the original stepwise model is $D_{app} - O$. This difference is a nearly unbiased estimate of the expected value of the external predictive discrimination of the process which generated D_{app}. In other words $D_{app} - O$ is an honest estimate of the internal validity, penalizing for over-fitting.**

Initially in Harrell's procedure the model M is developed using all of the original data, and Somer's D is recorded. In the next step I generate a bootstrap sample by drawing at random and with replacement from the original data. Next I fit a model M_1^* to this bootstrap sample, and record Somer's D, i.e. D_{boot}. Now obtain the Somer's D for M_1^* using the original data, i.e. D_{orig}. The optimism is defined to be $D_{boot} - D_{orig}$, here the optimism refers to M_1^*. If $D_{boot} - D_{orig}$ is < 5% this can be interpreted as meaning that the original model M is consistent in its performance, I do not see a degradation in predictive power when the original model is applied to the bootstrap data set. Although $D_{boot} - D_{orig}$ refers to M_1^*, I can say that the performance of M_1^* on the original data is at least comparable to that of M, i.e. I infer that M and M_1^* are the same model. In step 7 of Harrell's procedure I now run through steps 2 to 6 B times to obtain $M_1^*, M_2^*, ..., M_B^*$ and the associated D_{boot} and D_{orig}, which is denoted D_b^* and D_b^{*orig}. The quantity

$O = \frac{\sum_{b=1}^{B} D_b^* - D_b^{*orig}}{B}$ is the average optimism, $D_{app} - O$ gives a good estimate of the internal validity of the model, with O acting as a penalty term for over-fitting, large values for O mean I incur a high penalty for over-fitting. It is important to remember that a single value of Somer's D gives a measure of predictive power for a model, the difference in two values of Somer's D measures optimism or over-fitting. Harrell has implemented the validation procedure described in steps 1 to 7 in the Design library (Design Library Harrell Frank E. 2009b), (Design Library Harrell Frank E. 2009a).

6.4.0 Validating the CARE-HF Model

I shall look at an example of validating a model (the CARE-HF model) using Harrell's procedure. Note here the validation does not deal with optimism of the model fitting process, but from the final model alone. The model developed for the CARE-HF data has been described in Chapter 2, I will use GNU R and the Design library to validate the final model for the CARE-HF data. The variables in the final model for the CARE-HF data are shown in Table 6.1

	Transformation	Hazard ratio	95% CI	P-value
Predictors of overall outcome				
Mitral regurgitation	Log_e	1.71	1.38–2.12	0.0001
N-terminal pro-brain natriuretic peptide (pg/ml)	Log_e	1.31	1.17–1.47	0.0001
Systolic blood pressure (mmHg)	Linear	0.99	0.98–1.00	0.0698
Interventricular mechanical delay (ms)	Linear	1	0.99–1.01	0.7617
Aetiology (ischaemic) (yes/no)	Factor	1.89	1.45–2.46	0.0001
CRT (yes/no)	Factor	0.15	0.03–0.87	0.0347
Predictors of response to CRT				
Systolic blood pressure (mmHg)*CRT	Linear	1.02	1.00–1.03	0.0183
Interventricular mechanical delay (ms)*CRT	Linear	0.99	0.98–1.00	0.0084

Table 6.1 Predictors of outcome and response to CRT

Let us denote the variables in Table 1 as follows:

$\log_e$(Mitral regurgitation) x_1
$\log_e$(N-terminal pro-brain natriuretic peptide) x_2
Systolic blood pressure x_3
Interventricular mechanical delay x_4
Aetiology (ischaemic) x_5
CRT x_6
Systolic blood pressure*CRT $x_3 * x_6$
Interventricular mechanical delay*CRT $x_4 * x_6$
I will denote the primary event as p and the time to p as t.

To perform validation in GNU R using the Design library I use the following R code

```
>library(Design)
>setwd("C:/location of data file”)
>dd<-read.Table(file="myfile.csv",header=T,sep=",")
>attach(dd)
>f<-cph(formula=Surv(t,p)~x1+x2+(x4*x6)+(x3*x6)+x5,x=T,y=Y,surv=T)
>vl<-validate(f,B=200,dxy=T,pr=T)
```

I first load the Design library. Next we set the working directory and then read in the data file. I now make the data frame dd available through attach(). Next I fit the Cox Proportional Hazards model, justification for fitting the proportional hazards model even though there is some evidence that CRT violates this assumption is discussed in Chapter 2. f stores the result of the model fitting. Finally I validate the model using 200 bootstrap samples, dxy=T means that I want to use Somer's D, pr=T means print results for each of the 200 repetitions. The results of the validation procedure are shown in Table 6.2.

	index.orig	**training**	**test**	**optimism**	**index.corrected**	**n**
Dxy	-0.4090	-0.4198	-0.3982	-0.0216	-0.3874	200

Table 6.2 Validation of Final CARE-HF model Using Harrell's Design Library in GNU R

In Table 6.2 Somer's D (Dxy) is presented Dxy is the rank correlation between the predicted log hazard and the observed survival times. This is why we have the –ve values in Table 6.2, D_{app} = -0.41, the index corrected value for Somer's D (-0.3874) is a better estimate of the predictive power of the model, i.e. how well the model performs as a prognostic tool in the future. In terms of optimism I can interpret the value of -0.0216 from Table 6.2 as meaning that on average there is a difference of approximately 2% in the values of Somer's D between the original data and the 'new' data, so if the model where to be applied to a new set of patient data I would expect a

loss of predictive power of around 2%. As a rule of thumb an optimism of less than 5% is acceptable.

6.5.0 What Motivates Validation?

The phrase predictive power is broad description of the positive attributes that should be considered in regard to a prognostic model. Predictive power comprises two fundamental parts:

1. **Accuracy**
2. **Generalisabilty**

In (Justice *et al.* 1999) Justice defines accuracy as "The degree to which predicted outcomes match observed outcomes." Generalisabilty is defined as "Ability of a prognostic system to provide accurate predictions in a new sample of patients." (Justice *et al.* 1999). The aim of model validation is to assess whether the model is accurate and generalisable. Both accuracy and generalisabilty can be further broken down into the following parts:

Accuracy:

Calibration

Discrimination

Generalisabilty:

Reproducibility

Transportability

When considering accuracy, a calibration error occurs if the predicted probability of some event of interest is too high or too low. A discrimination error occurs if given that a patient has been assigned a risk score, they are incorrectly ranked on the basis

of individual risk. If patients are grouped based on their risk score, then the group comprising patients with a high score should have a high event rate, if a patient with a low risk score was allocated to the group with the high event rate, then a discrimination error has occurred. Similarly when considering generalisabilty, reproducibility refers to the accuracy of the prognostic model when applied to patients who were not in the original dataset used to develop the model, but are from the same population. If the prognostic model is accurate for patients from a similar but not identical population, or is accurate for data collected using methods that are different than those used to collect the original data; then the model can be said to possess reproducibility. It may appear that model validation is confined to assessing the validity of a model purely in statistical terms. Altman and Royston (Altman & Royston 2000) pose two questions of great importance:

1. With the available factors, is the model the best that can be found?
2. Does the model predict accurately enough for its purpose?

The above questions lead the authors of (Altman & Royston 2000) to suggest that validation be considered from both a statistical and a medical perspective. Altman and Royston (Altman & Royston 2000) supply the following definitions:

1. A *statistically validated model* is one which passes all appropriate statistical checks, including goodness-of-fit on the original data set and unbiased prediction on a new data set.
2. A *clinically validated model* is one which performs satisfactorily on a new data set according to context-dependent statistical criteria laid down for it.

I would concur with the view that it is necessary to distinguish between clinically and statistically validated models. In regard to Harrell's approach I believe that there is a

potential danger that to lose sight of the importance of clinical validation, Harrell's approach *appears* to concentrate on statistical validation. Researchers may be lulled into thinking that the validation methods suggested by Harrell are sufficient to produce a clinically useful prognostic model. As Altman and Royston point out, if the prognostic information is inherently weak, then a statistically valid model as defined in (Altman & Royston 2000) may be of limited use from a clinical perspective. The reader is strongly encouraged to consult (Altman & Royston 2000). The problem of model generalisation is of great interest to the present author, how far it is possible to produce general models is not clear, and the failure of a model to be general lies ultimately in the nature of the mathematical techniques used in model fitting. It is perhaps not un-reasonable to question anxiety over generalisabilty. Generalisabilty whilst desirable may be attainable to only a limited extent. I feel that this should be considered when carrying out statistical modelling. Clinicians want a prognostic model with good predictive power and ease of interpretation; it may well be that predictive power comes at the expense of ease of interpretation. This possibility will be discussed in the final chapter.

6.6.0 Summary

Validation is an important aspect of statistical modelling. Once I have obtained the 'final' model it is not enough to be content if this model fits the original data set well. Ideally I need to assess the performance of the model over new data, that is perform external validation. If it is not practical to perform external validation then I should apply some internal validation method such as data splitting, cross-validation or

bootstrap methods. If after performing external or internal validation I see deterioration in the predictive power of the model then I need to identify possible reasons, for example over-fitting (or under-fitting) misspecification of the functional form of the model. Another possible reason for poor performance of a model is missing data. If the original data set has significant missing data then this will influence the final model. I consider missing data and methods of imputing missing data in the next chapter. The model should be assessed as to whether it is clinically plausible; this is entirely separate from the issue of statistical validity.

CHAPTER 7 MISSING DATA AND IMPUTATION

- Missing can lead to a poor prognostic model
- Types of missing data MCAR, MAR and MNAR
- Missing data dealt with by imputation
- Correct imputation model is crucial
- Imputation should be used with caution

7.0.0 Introduction

If the data set contains variables for which values are missing then a model fitted to this data may not be reliable, missing data may lead to biased results. For example I could see how for variables with a large number of missing values, estimates of the regression coefficients could be distorted. In developing the prognostic model for the CARE-HF study (Richardson *et al.* 2007) it was found that mitral regurgitation was a strong predictor of the primary outcome. However mitral regurgitation was seen to have missing values (208 values were missing). Missing data may have a marked effect upon the variables that appear in the 'final' model. A variable may attain a spurious statistical significance due to missing values.

How do I treat the problem of missing data? This depends upon the reason for why the data is missing. Under certain circumstances the missing data will not lead to biased results. Unfortunately this is often not the case, and efforts must be undertaken to address the issue of missing data. One approach would be to remove cases where I have a high level of missing data. If I had a data set consisting of the variables y, x_1, x_2, and I wished to carry out a ordinary least squares regression of

y against x_1 and x_2, but I found that x_1 had a large number of missing values, then I might remove the pairs (y, x_1), where x_1 is missing. This method may lead to inflated variance and bias. The modern approach would be to supply the missing values. In the past the missing data problem tended to be ignored, imputation, the process of supplying or 'filling in' the missing values can be computationally intensive. Nowadays the computational power is available that makes imputation practical. An important question I must consider is why is the data missing? I shall now look at different types of missing data.

7.1.0 Types of Missing Data

If a group of patients have some measurement taken e.g. lung function, it is possible that some measurements may be missing due to failure of the measuring device or machine. In this situation I would assume that device or machine failure is a random event, the probability of missing data would be described as missing completely at random (MCAR). Other examples of situations where data would be described as missing completely at random are if for example someone was unable to complete a questionnaire due to common illness. Participants in a clinical study may move away from the area, they might die due to reasons unrelated to those specified within the study.

If the probability of missing data for a particular variable depends upon other observed variables then the missing data is said to be missing at random (MAR). If the probability of missing data for a particular variable depends on other observed variables and unobserved variables then data is said to be missing not at random (MNAR). MCAR MAR and MNAR are what is often called the missingness

mechanism, see (Buck 1960) and (Zhang 2003). If the missing data mechanism is MCAR or MAR then the missing data is said to be ignorable, the missingness mechanism does not need to be modelled, if however the missing data is MNAR the mssingness is said to be non-igonrable and is the most problematic of the missing data mechanisms. I need to determine why the data is missing, and once I have established that the data is not missing completely at random, then I should attempt to apply some suitable imputation method. It is possible to test the MCAR assumption, Little (Little 1988a) has developed a test based on the Chi squared distribution. However it is not possible to conclusively prove the data are MCAR. There is no test for the MAR assumption. For a detailed discussion of issue missing data and prognostic models see (Marshall 2007)

7.2.0 Dealing with Missing Data

There are numerous methods for dealing with missing data; I have mentioned one approach already, simply delete the missing data. It can be argued that this approach is not particularly satisfactory; as potentially useful information is being discarded (put another way, the sample size is reduced). I shall consider some of the methods available that allow missing data to be imputed. A very simple way of imputing data is to use the mean, missing values are replaced with the sample mean. For example in the CARE-HF data the variable mitral regurgitation has 208 missing values, if I impute these missing values by using the sample mean = 23.79 of the 605 non missing values for mitral regurgitation, then the sample mean for mitral regurgitation (n=813) with imputation = 23.79. Here I see that imputation using the sample mean has made no difference in the estimate of mean mitral regurgitation. What I do find however is that the standard deviations change, the standard deviation of the 605 non-

missing values = 14.94, whereas the standard deviation for the imputed data (n=813) = 12.88. This reduction in standard deviation is misleading in the sense that it is due to the fact that I have increased the sample size from 605 to 813, but I have seen no difference in the estimate for mean mitral regurgitation. However if I were to use the sample median, I have the sample median for the 605 non-missing values = 21.81 and the sample median for the imputed data = 23.79. Another possible approach is to impute the missing data by using some regression technique; I predict the missing values using the regression model. If I were to use ordinary least squares regression I am in effect doing the same thing as with using the sample mean, I am still confronted with the problem of producing a reduced standard deviation (or standard error) due to the increased sample size, but I will not gave gained any new information, i.e. I will not see an appreciable difference in the estimate of some population parameter based on the imputed data.

Table 7.1 briefly describes some of the common imputation methods

Method	**Comments**
Simple Mean Imputation, uses sample mean to impute missing values	Easy to perform, but may lead to distorted relationship between variable that has undergone imputation and other variables in dataset
Regression Imputation, use a regression model to generate missing values	Distribution of variable that has undergone imputation may be distorted, correlation with variable not included in the regression model may be suspect. If the regression model is not appropriate then imputed values are suspect.
Random Regression Imputation, as above but a random term is added to the imputed value generated by the regression model. Random term can be drawn from a normal distribution	Works well with categorical and continuous variables, again depends upon appropriate regression model.
Hot Deck Imputation, imputed value is selected at random from the non-missing cases	Method uses 'real' values, i.e. value is present in the data set.
Predictive Mean, a hot deck method that employs a regression model	Method is slightly more robust than the regression method
Last Value Carried Forward	Last known values carried forward to supply the missing data

Table 7.1 Imputation Methods

I shall now briefly review the basic ideas for the imputation methods that are implemented in SAS and GNU R. In SAS PROC MI (SAS Proc MI 2009) allows me to perform what is known as multiple imputation, see (Zhang 2003), (Rubin 1976), (Rubin 1996) and (Schafer & Olsen1998). The imputed data can then be analysed

using PROC MIANALYZE (SAS Proc Mianlyze 2009). In GNU R, Harrell's Design (Design Library Harrell Frank E. 2009b) library used in conjunction with Harrell's Hmisc library (Hmisc Library Harrell Frank E. 2009) allows me to perform imputation using the transcan and impute functions.

7.3.0 Multiple Imputation

So far my discussion of imputation has focused on trying to 'fill in' missing values for some variable, for each missing value of X I supply a single imputed value. Multiple imputation (Zhang 2003), (Rubin 1976), (Rubin 1996) does not supply a single imputed value, instead a set of possible values are considered. In multiple imputation I randomly sample from the existing data to generate this set of possible values. More formally multiple imputation can be described as follows:

1. Create k complete data sets by filling in all missing values k times, by drawing k times from the imputation model.
2. Analyse the k complete data sets, these data sets are regarded as real data.
3. Combine the results of the analysis of the k complete data sets to form the repeated or multiple imputation inference.

Earlier I looked at imputation using the sample mean, I noted problem with 'artificially' reduced standard deviation, multiple imputation overcomes the problem of reduced standard deviations or standard errors of estimates.

The imputation model is of fundamental importance, if I take the most simple case where the data set consists of one continuous variable X_1, then an example imputation model might be $X_1 \sim N(\mu, \sigma^2)$, the normal distribution model. If I have a data set that consisted of the continuous variables $X_1, X_2, X_3, ...X_p$, then I might use the multivariate normal model $X \sim N(m, \Sigma)$, where X is the vector

$(X_1, X_2, X_3, ...X_p)$ and m is a vector of means and Σ is the variance covariance matrix. For a mixture of binary and continuous variables I might use the conditional Gaussian (Horton & Kleinman 2009). In PROC MI it is possible to specify a 'customised' imputation model, for example I could specify that $X_1 = X_2$. I must consider that if I have specified a particular imputation model, then if someone else were to perform an analysis using the data set after I performed multiple imputation, there is a risk that this person may try fit a model different to that of my imputation model, for instance $X_1 = X_3 X_4$. It is advisable to use as many variables as possible when performing multiple imputation. For multiple imputation maximum likelihood estimates of parameters are obtained by using the EM algorithm (Dempster *et al.* 1977), (Gaetan & Yao 2003).

7.3.1 Imputation using Design and Hmisc

In GNU R the transcan function which is found in Harrell's Hmisc library performs both transformation and imputation for a variable. Results of applying the transcan and impute functions to the CARE-HF data are shown in Tables 7.2 to 7.8. Tables 7.9 and 7.10 present validation results for the final CARE-HF model with and without imputation. By default transcan uses single predicted expected value imputation, this is the case for the imputation performed here, it is possible to perform multiple imputation using transcan. If I want to perform multiple imputation using Design and Hmisc then the aregImpute function is a better choice; the results of performing multiple imputation using aregImpute are shown in Table 7.11. The main objective for the CARE-HF model was to determine possible treatment modifiers (interaction terms). For IVMD it is seen that for the imputed data coefficient is smaller and has undergone a sign change. The p-value for IVMD has decreased from 0.75900 to

0.06000, the interaction term IVMD*CRT is no longer statistically significant, this interaction term was borderline significant using the original data. Using imputed data the variable CRT is no longer significant, the p-value for systolic blood pressure has increased from 0.06959 to 0.34500, the interaction systolic blood pressure*CRT is no longer significant. Imputation has resulted in reductions in the coefficients for the variables mitral regurgitation, NT-pro-BNP and Ischaemia, all of these three variables remain highly significant. If I perform imputation using agreImpute then the interaction term systolic blood pressure*CRT is no longer significant, however the interaction term IVMD*CRT is just about significant at the 5% level. To reiterate, the main objective of the CARE-HF model was to identify possible treatment modifiers and not to produce a definitive prognostic tool, the treatment modifiers that were originally identified were admittedly weak. However it is interesting to note the effects of using different imputation methods, I see that the strong predictors have remained so, irrespective of the imputation method. The significance of CRT and the interaction terms differ notably dependent on whether imputation was performed or not. I would suggest that whilst this could be explained by missing data the fact that continuous variable had not been orthogonalization and binary variables not re-coded in the fashion described in Chapter 2 may have a considerable effect. I would recommend that orthogonalization should be carried out in situations where interactions between continuous and categorical variables are to be investigated. Where there is an appreciable level of missing data I would suggest that imputation should be performed, I would justify this based on the marked differences in the results for the model with and without imputation. However I would consider this in conjunction with orthogonalization

7.4.0 Summary

We should be aware that imputation is not without its dangers. Choosing an appropriate imputation model is crucial; if this imputation model is not appropriate then subsequent analysis will be flawed. One important point in regard to multiple imputation is that it has its origins in the problem of missing data in surveys, it has been suggested that this might be a limitation in terms of the efficiency of multiple imputation. Nielsen (Nielsen 2003) argues that Bayesian multiple imputation may not be efficient. For further discussion of some of the criticisms levelled at multiple imputation the reader is directed toward (Nielsen 2003) and (Rubin 2003). I would consider an investigation of predictive mean matching, see (Little Roderick 1988b) and (Heitjan & Little 1991), a useful exercise, Harrell's aregImpute function employs predictive mean matching. Predictive mean matching is an example of what is known as Hot Deck imputation (Altmayer 2009), Hot Deck imputation is one of the earliest imputation methods. Also in view of Nielsen's arguments (Nielsen 2003) an investigation of the methods used in Harrell's transcan and impute functions may be useful as Bayesian methods are an option for these functions.

Imputation is not a simple matter; a careful approach is needed when applying it. The literature relating to imputation is mathematically complex. The topic is a difficult one; even the basic definitions of MCAR, MAR and MNAR can be somewhat confusing when first encountered. In the next chapter I will look at the idea of the frailty model.

	coef	se(coef)	z	p
Mitral Regurgitation [a]	0.7590	0.1220	6.2000	0.0000
CRT	0.6140	0.5840	1.0500	0.2930
Mitral Regurgitation [a] *CRT	-0.3350	0.1820	-1.8300	0.0666
EVSI [a]	0.4162	0.1750	2.3761	0.0175
CRT	-0.4752	1.4120	-0.3366	0.7360
ESVI [a]*CRT	-0.0013	0.2920	-0.0044	0.9965
Ischaemic	0.5220	0.1350	3.8800	0.0001
CRT	-0.7310	0.1580	-4.6200	0.0000
Ischaemic*CRT	0.4010	0.2120	1.8900	0.0584
Ejection Fraction [a]	-0.9730	0.2840	-3.4280	0.0006
CRT	-1.1470	1.4470	-0.7930	0.4280
Ejection Fraction [a] *CRT	0.2180	0.4550	0.4790	0.6320
Age	0.0237	0.0073	3.2620	0.0011
CRT	-0.1452	0.7284	-0.1990	0.8420
Age*CRT	-0.0051	0.0108	-0.4690	0.6390
Systolic Blood Pressure	-0.0130	0.0040	-3.2600	0.0011
CRT	-1.9373	0.7484	-2.5900	0.0096
Systolic Blood Pressure*CRT	0.0126	0.0064	1.9700	0.0489
Glomerular Filtration Rate	-0.0129	0.0037	-3.5070	0.0005
CRT	-0.2967	0.3495	-0.8490	0.3960
Glomerular Filtration Rate*CRT	-0.0032	0.0057	-0.5520	0.5810
NT-pro-BNP [a]	0.3887	0.0589	6.5990	0.0000
CRT	-1.1054	0.7271	-1.5200	0.1280
NT-pro-BNP [a]*CRT	0.0785	0.0905	0.8670	0.3860
IVMD	-0.0077	0.0026	-2.9950	0.0028
CRT	-0.0827	0.1985	-0.4160	0.6770
IVMD*CRT	-0.0077	0.0039	-1.9840	0.0473

a = log transformed, * denotes an interaction

Table 7.2 Univariate Models For Each Potential Predictor (without imputation)

	coef	se(coef)	z	p
Mitral Regurgitation [a]	0.5299	0.0833	6.3639	0.0000
CRT	-0.0036	0.3947	-0.0092	0.9930
Mitral Regurgitation [a] *CRT	-0.1692	0.1253	-1.3506	0.1770
EVSI [a]	0.4225	0.1610	2.6228	0.0087
CRT	-0.0703	1.1980	-0.0587	0.9530
ESVI [a]*CRT	-0.0807	0.2480	-0.3252	0.7450
Ischaemic	0.5220	0.1350	3.8800	0.0001
CRT	-0.7310	0.1580	-4.6200	0.0000
Ischaemic*CRT	0.4010	0.2120	1.8900	0.0583
Ejection Fraction [a]	-0.9110	0.2490	-3.6630	0.0003
CRT	-1.3420	1.1680	-1.1490	0.2510
Ejection Fraction [a] *CRT	0.2730	0.3670	0.7430	0.4580
Age	0.0237	0.0073	3.2620	0.0011
CRT	-0.1452	0.7284	-0.1990	0.8420
Age*CRT	-0.0051	0.0108	-0.4690	0.6388
Systolic Blood Pressure	-0.0125	0.0040	-3.1600	0.0016
CRT	-1.6131	0.7302	-2.2100	0.0272
Systolic Blood Pressure*CRT	0.0099	0.0062	1.5800	0.1130
Glomerular Filtration Rate	-0.0146	0.0035	-4.1500	0.0000
CRT	-0.3004	0.3237	-0.9280	0.3530
Glomerular Filtration Rate*CRT	-0.0032	0.0055	-0.5940	0.5520
NT-pro-BNP [a]	0.3741	0.0564	6.6300	0.0000
CRT	-0.4602	0.6637	-0.6930	0.4880
NT-pro-BNP [a]*CRT	-0.0104	0.0832	-0.1250	0.9000
IVMD	-0.0058	0.0021	-2.7300	0.0064
CRT	-0.2558	0.1676	-1.5300	0.1270
IVMD*CRT	-0.0046	0.0031	-1.4800	0.1400

a = log transformed, * denotes an interaction

Table 7.3 Univariate Models For Each Potential Predictor (with imputation)

Mitral Regurgitation [a]	Obs	Events	Model L.R.	d.f.	P	Score	Score P	R2	Imputed
	605	289	66.8400	3.0000	0.0000	67.4200	0.0000	0.1100	N
	813	383	86.3100	3.0000	0.0000	82.5700	0.0000	0.1000	Y
EVSI [a]	732	349	28.9300	3.0000	0.0000	29.4700	0.0000	0.0400	N
	813	383	30.4900	3.0000	0.0000	31.3700	0.0000	0.0400	Y
Ischaemic	812	383	67.8300	3.0000	0.0000	65.9500	0.0000	0.0800	N
	813	383	67.8100	3.0000	0.0000	65.9200	0.0000	0.0800	Y
Ejection Fraction [a]	745	357	33.4800	3.0000	0.0000	35.4000	0.0000	0.0400	N
	813	383	38.4900	3.0000	0.0000	41.1300	0.0000	0.0500	Y
Age	813	383	36.8700	3.0000	0.0000	37.7300	0.0000	0.0400	N
	813	383	36.8700	3.0000	0.0000	37.7300	0.0000	0.0400	Y
Systolic Blood Pressure	803	378	31.8900	3.0000	0.0000	34.0500	0.0000	0.0400	N
	813	383	30.9900	3.0000	0.0000	32.9200	0.0000	0.0400	Y
Glomerular Filtration Rate	739	338	45.9800	3.0000	0.0000	43.8600	0.0000	0.0600	N
	813	383	59.2000	3.0000	0.0000	56.1100	0.0000	0.0700	Y
NT-pro-BNP [a]	732	346	109.3300	3.0000	0.0000	105.0800	0.0000	0.1400	N
	813	383	102.2200	3.0000	0.0000	101.0200	0.0000	0.1200	Y
IVMD	735	346	52.3500	3.0000	0.0000	49.2400	0.0000	0.0700	N
	813	383	45.9500	3.0000	0.0000	45.1000	0.0000	0.0600	Y

a = log transformed

Table 7.4 Fit Statistics for Univariate Models

	coef	se(coef)	z	p
Mitral Regurgitation [a]	0.5381	0.1088	4.9470	0.0000
IVMD	0.0010	0.0034	0.3070	0.7590
CRT	-1.8753	0.8876	-2.1130	0.0346
Systolic Blood Pressure	-0.0087	0.0048	-1.8150	0.0695
NT-pro-BNP [a]	0.2720	0.0591	4.6010	0.0000
Ischaemic	0.6345	0.1349	4.7050	0.0000
IVMD*CRT	-0.0131	0.0050	-2.6390	0.0083
Systolic Blood Pressure*CRT	0.0172	0.0073	2.3600	0.0183

a = log transformed

Table 7.5 Coefficients For Final Model (without imputation)

Obs	Events	Model L.R.	d.f.	P	Score	Score P	R2
526	249	130.1900	8.0000	0.0000	121.0200	0.0000	0.2200

Table 7.6 Fit Statistics For Final Model (without imputation)

	coef	se(coef)	z	p
Mitral Regurgitation [a]	0.3131	0.0800	4.1100	0.0000
IVMD	-0.0046	0.0025	-1.8810	0.0600
CRT	-1.1993	0.7073	-1.6960	0.0900
Systolic Blood Pressure	-0.0035	0.0038	-0.9440	0.3450
NT-pro-BNP [a]	0.2362	0.0513	4.6070	0.0000
Ischaemic	0.5280	0.1094	4.8260	0.0000
IVMD*CRT	-0.0057	0.0034	-1.6550	0.0979
Systolic Blood Pressure*CRT	0.0075	0.0058	1.2950	0.1950

Table 7.7 Coefficients For Final Model (with imputation)

Obs	Events	Model L.R.	d.f.	P	Score	Score P	R2
813	383	174.7200	8.0000	0.0000	165.3600	0.0000	0.1900

Table 7.8 Fit Statistics For Final Model (with imputation)

	index.orig	training	test	optimism	index.corrected	n
Dxy	-0.4090	-0.4233	-0.3984	-0.0249	-0.3841	200

Table 7.9 Validation Results For Final Model (without imputation)

	index.orig	training	test	optimism	index.corrected	n
Dxy	-0.3919	-0.3984	-0.3855	-0.0130	-0.3789	200

Table 7.10 Validation Results For Final Model (with imputation)

	coef	se(coef)	z	p
Mitral Regurgitation [a]	0.3799	0.0868	4.3800	0.0000
IVMD	-0.0046	0.0028	-1.6500	0.0984
CRT	-0.9963	0.7288	-1.3700	0.1720
Systolic Blood Pressure	-0.0046	0.0038	-1.2100	0.2280
NT-pro-BNP [a]	0.2939	0.0471	6.2400	0.0000
Ischaemic	0.5539	0.1094	5.0600	0.0000
IVMD*CRT	-0.0085	0.0043	-2.0000	0.0455
Systolic Blood Pressure*CRT	0.0070	0.0059	1.1900	0.2350

a = log transformed

Table 7.11 Coefficients For Final Model (with multiple imputation (5 imputations))

Chapter 8 Frailty Models

- Individual and group heterogeneity modelled by random effects
- For time to event data the equivalent of the random effects model is the frailty model
- The frailty model is an extension of the proportional hazards model
- After fitting a frailty model to the CARE-HF data it is found that the conventional proportional hazards model is adequate

8.0.0 Introduction

A natural extension of the prognostic model developed for the CARE-HF data (see chapter 2) would be to consider a frailty model. For the CARE-HF data information was recorded on treatment centre. To help understand the idea of a frailty model it might be helpful to first look at the possibly more familiar random effects model. In the regression model

$Y_{jk} = \beta_0 + \beta_1 X_k + \varepsilon$, where Y_{jk} is a measurement on the j^{th} subject (patient) at time X_k, and where $\varepsilon \sim N(0, \sigma^2)$, the population parameters β_0 and β_1 are estimated, the intercept and slope of the line, to obtain $\hat{Y}_{jk} = \hat{\beta}_0 + \hat{\beta}_1 X_k$ or $Y_{jk} = \hat{\beta}_0 + \hat{\beta}_1 X_k + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. I have repeated measures on each subject; also I have treated the data as being homogeneous in the sense that the slope and intercept are the same for each subject. What if the data are not homogeneous?

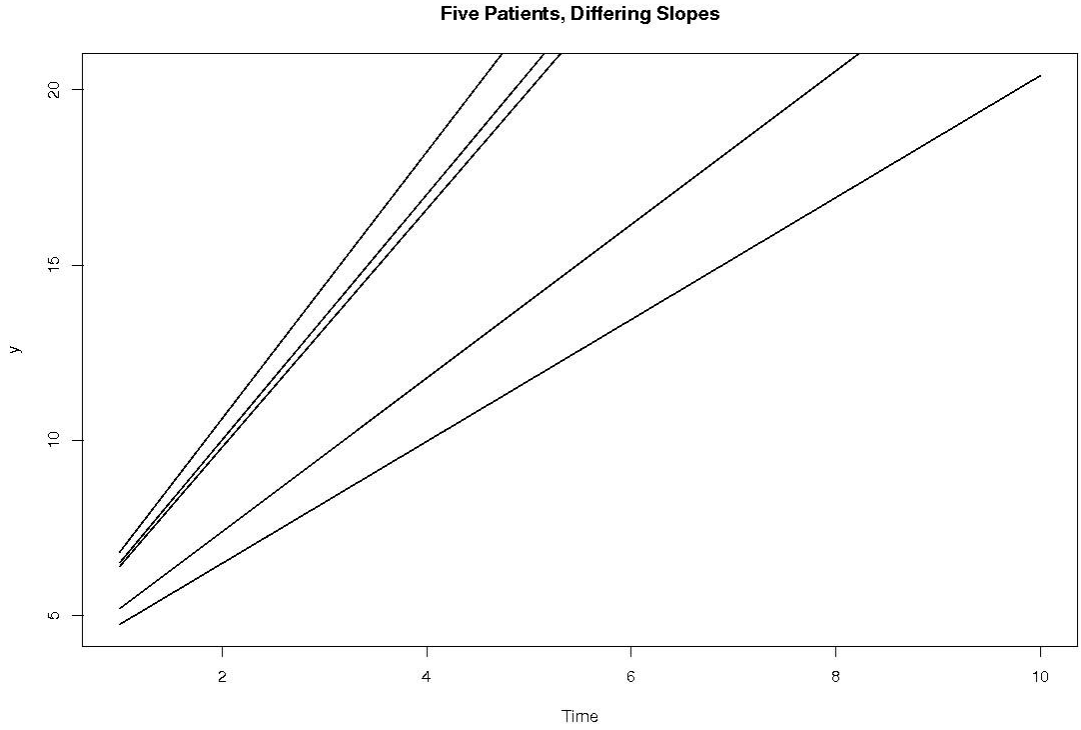


Figure 8.1 measurement on 5 hypothetical patients differing slopes

The figure 8.1 above illustrates a hypothetical situation for five patients on whom ten measurements Y have been taken at different times X_k, it appears that the intercept for each patient is the same, but that the slope varies from patient to patient. I need to develop a model that takes into account the varying slopes. Assuming that the variation in the slopes is random, the model $Y_{jk} = \beta_0 + (\alpha_j + \beta_1)X_k$ incorporates the random slopes through the term α_j .The term α_j is known as a random effect.

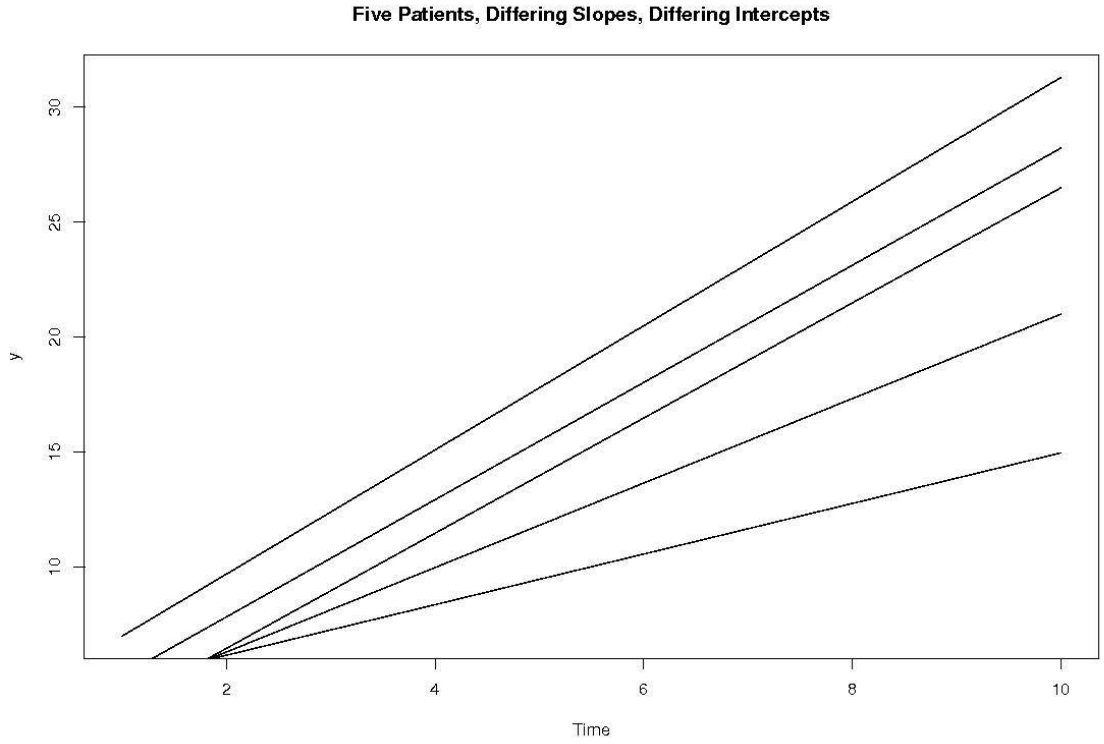


Figure 8.2 measurements on 5 hypothetical patients, differing slopes and intercepts

The figure 8.2 shows a situation where both the intercept and the slope vary from patient to patient, again I need to incorporate the varying slope and intercept into the model. The model $Y_{jk} = \beta_0 + \zeta_j + (\alpha_j + \beta_1)X_k$ now contains an additional random effect ζ_j, the random intercept; such a model is known as a mixed model. It is important to point out that I am not interested in obtaining numerical estimates for α_j and ζ_j, I am concerned with whether or not their inclusion improves the model fit. I could consider a situation in which I have data on a number of patients who have received treatment at several different hospitals or clinics, for each patient I would have a repeated measure Y_{jk} taken at time X_k, also let C_m indicate at which of the m hospitals or clinics the patient received their treatment. I might find that the slope, intercept or both vary depending upon which of the hospitals or clinics the patient attended. In this case the model should include α_j and/or ζ_j, random effects for the intercept and slope.

8.1.0 Frailty

The idea of frailty is another way of incorporating random effects and heterogeneity into a model for time to event data (survival model). In most biomedical and epidemiological applications the time to event data is assumed to be homogeneous, in reality there may be sources of unobserved heterogeneity within the data. For example, if $\tilde{x}$ is a vector of independent variables (co-variants), it is quite possible that X_{unob}, some powerful predictor of Y, is missing for whatever reason. It is not practical to include all possible covariates, such as when the number of events within a particular stratum is very small, or it may be that the particular co-variate has yet to be identified.

In a clinical trial, one important potential source of heterogeneity is the treatment centre. Section 3.2 of ICH E9 (ICH E9. 1999), (ICH HARMONISED TRIPARTITE GUIDELINE 1998) which addresses multicentre trials, places great emphasis on a proper treatment of centre effects and states:

"Up to this point the discussion of multicentre trials has been based on the use of fixed effect models. Mixed models may also be used to explore the heterogeneity of the treatment effect. These models consider centre and treatment-by-centre effects to be random, and are especially relevant when the number of sites is large." Use of frailty models would seem to be in accord with the guidelines laid down in ICH E9 although at present their use is not advocated.

Taking the proportional hazards model $h_1(y) = h_0(y)e^{\tilde{\beta}\tilde{x}}$, the hazard for each subject will be different and determined by X_{unob}. How can I include unobserved co-variates

in the proportional hazards model? The answer lies in the idea of frailty. Frailty could be described as accident proneness, or in terms of the force of mortality upon a certain subject (force of mortality is the hazard function $h(y)$).

Vaupel (Vaupel *et al.* 1979) defines frailty in the following way, let $\mu_i(x, y, z)$ be the force of mortality for an individual in population group i, at exact age x, at time y with frailty z, then $\frac{\mu_i(x, y, z)}{\mu_i(x, y, z')} = \frac{z}{z'}$

Now $z' = 1$ describes a 'standard individual', so we get $\mu_1(x, y, z) = z\mu_i(x, y, 1)$. An individual with a frailty of 3 is 3 times as likely to die or experience the event of interest as the standard individual. Following Vaupel's notation I write $\mu_i(x, y, z)$ as $\mu(z)$, $\mu_i(x, y, 1)$ as $\mu(1)$ or μ. So I have $\mu(z) = z\mu$. I could apply this idea to the proportional hazards model to arrive at $\frac{\mu_1(y, z)}{\mu_0(y)} = z\exp(\widetilde{\beta x})$. Rearranging the above formula gives $\mu_1(y, z) = \mu_0(y)z\exp(\widetilde{\beta x})$. The above is an example of a univariable frailty model; this frailty model is an extension of Cox Proportional Hazards model . The frailty model is the equivalent of the random effects model for time to event data. It must remembered that z is a random variable, also I must have $z \geq 0$, this dictates the choice of distribution for z.

Typical choices for the distribution of z include the Gamma distribution $f(z, \lambda) = \frac{z^{\varphi-1}\lambda^{\varphi}e^{-z\lambda}}{\Gamma(\varphi)}$.Where λ is scale parameter, φ is a shape parameter and $\Gamma(\lambda)$ is the Gamma function $\Gamma(\lambda) = \int_0^{\infty} u^{\lambda-1}e^{-u}du$. The Gamma distribution is a

logical choice for the distribution of z ; as z is non-negative this makes the Gamma distribution a sensible choice. Vaupel (Vaupel *et al.* 1979) states that frailty was assumed to follow the Gamma distribution because the distribution is "analytically tractable and readily computable". The Gamma distribution is flexible in the sense that, as φ varies, the distribution can take on different shapes. Also in Vaupel (Vaupel *et al.* 1979) describes two convenient mathematical results that arise from the assumption that frailty follows the Gamma distribution. I see that if $z < 1$, then the hazard for an individual will be reduced, and if $z > 1$, then the hazard is increased.

The important point is that, in the frailty model, the hazard for an individual is determined by both observed and unobserved factors. The following papers (Wienke 2003), (Manton *et al.* 1986), (Hougaard 1991), (Hougaard 1984) and (Perperoglou *et al.* 2007) are highly informative and contain material detailing the motivation and development of frailty models along with discussion on the issue of the distribution of the frailty. Including frailty in a prognostic survival model seems to be a very natural and highly appealing thing to do.

In recent years, faster CPUs have meant that some of the previous difficulties (relating to numerical methods) encountered when trying to fit frailty models have been overcome. Consequently, it is now quite possible to fit a frailty model in situations where previously this may have been difficult and we no longer have to simply ignore centre effects.

8.2.0 Fitting a Frailty Model to the CARE-HF data

I shall now proceed to fit a univariable gamma frailty model to the CARE HF data (Richardson *et al.* 2007) whilst at the same time applying elements of Harrell et al.'s (1996) approach. Earlier the following covariates where identified as being potential predictors of outcome and response to CRT:

- Mitral Regurgitation (MR)
- Interventricular Mechanical Delay (IVMD)
- End-systolic volume index (ESVI)
- Glomerular Filtration Rate (GFR)
- Systolic Blood Pressure (SBP)
- Ejection Fraction (EF)
- N-terminal pro-brain natriuretic peptide (NT-pro-BNP)
- Age
- Aetiology (Ischaemic)

As before I start by fitting a proportional hazards model for each of the potential predictors identified above (univariable analysis), using Mitral Regurgitation as an example we would fit the model $MR + (MR * CRT) + CRT$ where $MR * CRT$ is an interaction term. We assume also that the transformations applied in Chapter 3 are still used, so we would consider $\log_e(MR) + (\log_e(MR) * CRT) + CRT$. Now in addition I shall include a frailty term, the frailty term is assumed to follow the gamma distribution; this extended Cox Proportional Hazards model is a gamma frailty model.

For the CARE-HF study the treatment centre that each patient attended was recorded in the form of the variable SiteNum (site number), for the CARE-HF study there were 82 centres across Europe. Centre effects are modeled using the idea of grouped frailty, for example patients who received treatment at the same hospital would be regarded as sharing a common frailty. Centre effects are of interest due varying clinical skills, case-mix, technology, funding and so on. I shall model site number as a grouped frailty term, i.e. each treatment centre represents a group of patients; frailty can be also be modeled at an individual level, an individual patient characteristic could be treated as a frailty term.

Univariate models are produced for each of the other potential predictors; I then include significant (5% level) covariates and interaction terms from these univariable models as candidates in the final model. The models were fitted using coxph (R survival package Terry Therneau 2009) from the recommended base survival package in GNU R version 2.7.2 (R Foundation for Statistical Computing 2009). In Tables 8.1-8.9, coefficients are presented for each of the univariable models. Note that, with coxph, automated stepdown or stepwise selection is not possible. Table 8.11 shows the final conventional Cox Proportional Hazards model presented in (Richardson *et al.* 2007) (see Chapter 2 for a full discussion of this model).

The final frailty model shown in Table 8.10 is obtained in the following way: all co-variates that are statistically significant (5% level) in the univariable analysis are considered as candidates for inclusion in the final model; a non-stepwise backward selection procedure is then applied resulting in the final (frailty) model. It can be seen from Table 8.10 that SiteNum is not significant; this suggests that the conventional

Cox Proportional Hazards model may be adequate, i.e. the data not observed to be heterogeneous with respect to treatment centre. However it can be argued that even if the frailty term is not statistically significant, it should be retained, i.e. we adopt the frailty model. The overheads in terms of model complexity and computational resources are not so great that we would abandon the frailty model in favour of the conventional Cox Proportional Hazards model. It may in fact be natural and appropriate as far as the design of a model is concerned to include a frailty term.

Comparing Tables 8.10 and 8.11 it is seen that the final models are similar. The likelihood ratio test for the frailty model gives a slightly larger value than that for the Cox Proportional Hazards model, however this result is not statistically significant. The confidence interval and p-values produced for both models are consistent. If heterogeneity had been present in the data, and I was to fit a conventional Cox Proportional Hazards model I am liable to obtain confidence intervals that are too narrow and p-values that are too small. The frailty models I have considered are relatively simple, for example I have not attempted to fit a frailty model where some of the co-variates required transformation via cubic splines or fractional polynomials. Validation of the frailty model presented in Table 8.10 was not performed as both Therneau's survival package and Harrell's Design package do not have the facility to validate frailty models. This is a drawback I hope that at some point in the future it will be possible to routinely validate frailty models in GNU R. As far as I am aware the situation is no different in SAS, in fact it is rather difficult to even produce frailty models easily and efficiently in SAS.

8.2.1 Tables 8.1-8.9 Univariable Frailty Models For Each Potential Predictor (without imputation)

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
Mitral Regurgitation [a]	2.1300	1.6820	2.7200	p<0.0001
CRT	1.8450	0.5870	5.8000	0.2900
frailty(SiteNum)				0.2400
Mitral Regurgitation [a] ***CRT**	0.7160	0.5010	1.0200	0.0670

Likelihood ratio test = 67.9 on 3.49 df, p<0.0001
a = log transformed, * denotes an interaction
Table 8.1 Univariate Analysis Mitral Regurgitation (MR) n=605 (208 observations deleted due to missingness)

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
EVSI [a]	1.5250	1.0715	2.1700	0.0190
CRT	0.5490	0.0326	9.2200	0.6800
frailty(SiteNum)				0.1600
ESVI [a] ***CRT**	1.0220	0.5701	1.8300	0.9400

Likelihood ratio test = 62 on 17.0 df, p<0.0001
a = log transformed, * denotes an interaction
Table 8.2 Univariate Analysis End-systolic volume index (ESVI) n=732 (81 observations deleted due to missingness)

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
Ischaemic	1.7070	1.3030	2.2360	p<0.0001
CRT	0.4770	0.3490	0.6520	p<0.0001
frailty(SiteNum)				0.1900
Ischaemic*CRT	1.4980	0.9850	2.2780	0.0590

Likelihood ratio test = 97.6 on 15.7 df, p<0.0001

* denotes an interaction

Table 8.3 Univariate Analysis Aetiology (Ischaemic) n=812 (1 observation deleted due to missingness)

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
Ejection Fraction [a]	0.3820	0.2176	0.6700	0.0008
CRT	0.3310	0.0191	5.7300	0.4500
frailty(SiteNum)				0.3200
Ejection Fraction [a] ***CRT**	1.2260	0.5001	3.0000	0.6600

Likelihood ratio test = 45.2 on 8.3 df, p<0.0001

a = log transformed, * denotes an interaction

Table 8.4 Univariate Analysis Ejection Fraction (EF) n=745 (68 observations deleted due to missingness)

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
Age	1.0240	1.0090	1.0400	0.0014
CRT	0.8500	0.2030	3.5600	0.8200
frailty(SiteNum)				0.3400
Age*CRT	0.9950	0.9740	1.0200	0.6600

Likelihood ratio test = 45.9 on 7.09 df, p<0.0001

a = log transformed, * denotes an interaction

Table 8.5 Univariate Analysis Age n= 813

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
Systolic Blood Pressure	0.9870	0.9795	0.9950	0.0012
CRT	0.1440	0.0332	0.6260	0.0098
frailty(SiteNum)				0.2900
Systolic Blood Pressure*CRT	1.0130	1.0000	1.0250	0.0490

Likelihood ratio test = 36.2 on 4.77 df, p<0.0001

a = log transformed, * denotes an interaction

Table 8.6 Univariate Analysis Systolic Blood Pressure (SBP) n=803 (10 observations deleted due to missingness)

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
Glomerular Filtration Rate	0.9870	0.9800	0.9940	p<0.0001
CRT	0.7420	0.3740	1.4730	0.3900
frailty(SiteNum)				0.2500
Glomerular Filtration Rate*CRT	0.9970	0.9860	1.0080	0.5800

Likelihood ratio test = 47.1 on 3.52 df, p<0.0001

a = log transformed, * denotes an interaction

Table 8.7 Univariate Analysis Glomerular Filtration Rate (GFR) n=739 (74 observations deleted due to missingness)

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
NT-pro-BNP [a]	1.4750	1.3142	1.6600	p<0.0001
CRT	0.3310	0.0796	1.3800	0.1300
frailty(SiteNum)				0.8900
NT-pro-BNP [a]*CRT	1.0820	0.9058	1.2900	0.3900

Likelihood ratio test = 109 on 3 df, p<0.0001

a = log transformed, * denotes an interaction

Table 8.8 Univariate Analysis N-terminal pro-brain natriuretic peptide (NT-pro-BNP) n=732 (81 observations deleted due to missingness)

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
IVMD	0.9920	0.9870	0.9970	0.0023
CRT	0.9250	0.6240	1.3690	0.7000
frailty(SiteNum)				0.2000
IVMD*CRT	0.9920	0.9840	1.0000	0.0400

Likelihood ratio test = 77.5 on 13.6 df, p<0.0001

a = log transformed, * denotes an interaction

Table 8.9 Univariate Analysis Interventricular Mechanical Delay (IVMD) n=735 (78 observations deleted due to missingness)

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
Mitral Regurgitation [a]	1.7160	1.3857	2.1240	p<0.001
IVMD	1.0010	0.9944	1.0080	0.7700
CRT	0.1510	0.0265	0.8670	0.0340
frailty(SiteNum)				0.2700
Systolic Blood Pressure	0.9914	0.9821	1.0007	0.0720
NT-pro-BNP [a]	1.3140	1.1699	1.4759	p<0.001
Ischaemic	1.8887	1.4491	2.4618	p<0.001
IVMD*CRT	0.9870	0.9774	0.9966	0.0084
Systolic Blood Pressure*CRT	1.0174	1.0030	1.0321	0.0180

Likelihood ratio test = 133 on 9.07 df, p<0.0001

a = log transformed, * denotes an interaction

Table 8.10 Final model n=526 (287 observations deleted due to missingness)

	Hazard Ratio	CI lower 95 %	CI upper 95 %	p
Mitral Regurgitation [a]	1.7128	1.3839	2.1199	p<0.001
IVMD	1.0010	0.9945	1.0080	0.7600
CRT	0.1533	0.0269	0.8733	0.0350
Systolic Blood Pressure	0.9914	0.9821	1.0010	0.0700
NT-pro-BNP [a]	1.3126	1.1690	1.4739	p<0.001
Ischaemic	1.8868	1.4486	2.4576	P<0.001
IVMD*CRT	0.9870	0.9774	0.9966	0.0083
Systolic Blood Pressure*CRT	1.0173	1.0029	1.0320	0.0180

Likelihood ratio test = 130 on 8 df, p<0.0001

a = log transformed, * denotes an interaction

Table 8.11 Final (non frailty) model

CHAPTER 9 CONCLUSION

- Model selection is a complex problem
- Quality of prognostic models related to limitations in methodology
- Techniques such as MDA may offer possible alternative approaches
- Ideas such as over-fitting, optimism should be introduced in elementary courses on statistical modelling
- More user friendly software might help researchers to produce better models

9.0.0 Introduction

In this work I have been concerned with producing a good quality prognostic model for the CARE-HF data (Richardson *et al.* 2007). The prognostic model developed for the CARE-HF data represents a significant real world example of a prognostic model, I am not aware that the model has been made use of in practice. The prognostic model for the CARE-HF data indicates that all patients are likely to benefit from cardiac resynchronisation therapy, i.e. the treatment modifiers identified in the model are weak. One way of validating the model developed for the CARE-HF data would be to apply it to new data. Apart from the COMPANION study (Bristow *et al.* 2004) there have been no investigations comparable to the CARE-HF study. Unfortunately individual patient data from the COMPANION study although requested has not been made available. It is unlikely that further investigation of CRT will be undertaken as its benefit has been established. The difficulties and problems encountered when producing this model are likely to be experienced by other researchers when they make efforts to deal issues such as functional form, over-fitting, optimism and validation. In producing the model for the CARE-HF data I found that when one

attempts to employ a strategy such as the one suggested by Harrell and colleagues (Harrell *et al.* 1996) one ends up having to consider the fundamental problem of model fitting. The topic of model selection is a deep one; I have great admiration for the skill and insight displayed by researchers into this problem.

I have come to an appreciation of the complexity surrounding the problem of developing a good prognostic model. This work it is hoped has served as an accessible guide to some of the main methods that feature in the process of fitting a prognostic model, (or a model in general). Implementing an approach such as Harrell's is not a trivial task. I believe that I have identified some important limitations in Harrell's strategy. I shall now present a brief summary of the material covered in the course of this work and indicate important points that have arisen.

9.1.0 Summary of Main Topics

Chapter 1 contained an introductory discussion a definition of a prognostic model was given, the problem of over-fitting was introduced along with the idea of optimism. In Chapter 2 the prognostic model developed for the CARE-HF data was described and discussed. The development of this model was in itself a substantial piece of work. Absolute risk estimates and risk score where discussed in chapter 3, I presented a risk score calculator based on the prognostic model for the CARE-HF data. The problem of functional form was investigated in chapter 4. Use of cubic splines and fractional polynomials was discussed. In chapter 5 model fit was considered, the AIC was described in some detail. Over-fitting and optimism were discussed in further detail in chapter 6. Validation methods were also considered. Missing data and imputation were discussed in chapter 7. Chapter 8 introduced the idea of a frailty model.

This work has provided me with a great many questions and future areas of investigation. In chapter 1 of this work I stated that the whole question of

generalizabilty is a complex one. The following questions were then asked. Should we expect to achieve more general results in the physical sciences? Do biomedical applications present us with special problems? The answer to both of these questions is probably no. Over-fitting is a problem for researchers working in the fields of Physics, Astronomy and other physical sciences. Over-fitting is also a problem in Ecological, Economic and Financial models (Ginzburg & Jensen 2004). The prognostic model described in chapter 2 of this work prompts me to consider a practical question that researchers may have to consider in regard to choice of software. An implementation of Harrell's design library does exist for SAS; however this is an old version, development is focused on the S-Plus and R versions. Given that SAS is a widely used system an up-to-date and user friendly version for SAS would be of great value. There may be many researchers who for a variety of reasons may not be able to adopt R or S-Plus. Hmisc also seems to suffer from a lack of up-to-date versions that could easily be installed on a recent version of SAS.

In the discussion of cubic splines in chapter 4 it might be useful to reflect on Harrell's use of the restricted cubic spline (Harrell *et al.* 1996), (Herndon & Harrell 1990). What clinical/biological evidence there is to support this particular choice for the functional form of the model? Use of cubic splines may improve the fit of the model (on the original data), and when considering model validation goodness of fit is a basic criterion. As Altman and Royston point out in (Altman & Royston 2000) a statistically valid model may be clinically invalid. A choice of functional form that improves model fit and so leads to a statistically valid model, may not lead to a clinically valid model. The biological plausibility of the model is a matter for the medical expert to consider. Harrell's strategy has been central element of many of the discussions in the work, what general remarks would I make about Harrell et al.'s

approach? I am of the view that adopting Harrell's recommendations for avoiding over-fitting should certainly be included as part of the process of fitting a prognostic model. Harrell encourages inter-disciplinary collaboration, clinicians should be consulted by the statistician throughout the modelling process, this is vital if the model is to be a sensible. Harrell's approach allows the researcher to determine if there is a risk of over-fitting by use of the inequality $\frac{N_E}{p} < 10$. The extent of over-fitting is gauged via an estimation of the optimism. What in built mechanism exists within Harrell's approach that will minimise the risk of over-fitting? This question could be answered by noting that model selection based on the AIC or BIC is implemented as part of Harrell's software. However using Harrell's approach it is still quite easy to produce a model that is over-fitted. In chapter 8 of this work frailty models were considered, one of the limitations of Harrell's approach is that it does not encompass frailty models. I believe this to be a significant omission. What alternative approach could be adopted that might enable the researcher to produce reliable and accurate prognostics models? I will now outline some areas that I have found to offer potentially useful alternative methods to model fitting. In no way do I mean to suggest that they are better than the strategy devised by Harrell, but may offer fruitful areas for further investigation.

9.2.0 Alternative Modelling Techniques

Throughout this work the modelling techniques discussed have been what could be described as traditional, i.e. the Cox proportional hazards model. Are there alternatives to the traditional approaches to modelling that could offer better results in regard to the problem of over-fitting?

9.2.1 Data Modelling and Algorithmic Modelling

It is not uncommon for a researcher new to statistical modelling to assume that the more variables that are included in a model the better. This is on the surface a quite reasonable assumption, as the number of variables is in some way equated to 'information'. The more variables we have in the model, the better the description of reality provided by the model. It can be difficult for someone to appreciate the concept of parsimony in statistical modelling. In some introductory courses on statistics the idea of parsimony is mentioned as an important feature of a model, but the reason for its importance may not be clearly elucidated. It might be argued that if researchers new to statistical modelling do not appreciate the idea of a parsimonious model they may be likely to get into difficulty with over-fitting. The conventional view is that parsimony is a necessary and desirable characteristic for a model, one would expect a simple parsimonious model to be more easily interpreted than a complex model containing a large number of variables. There is an alternative view of parsimony; it could be argued that by seeking to produce a model based on Occam's razor unrealistic restrictions have been imposed. Real world situations such as those presented in medicine involve complex mechanisms; therefore the model may be extremely complex. Breiman argues in (Breiman 2001) that instead of aiming to

minimise the dimension of a model it should be increased. Breiman propounds the idea that so far as predictive accuracy is concerned the best model is the most complex one; in fact so complex that it may defy interpretation.

Breiman describes two approaches toward statistical modelling

- Data Modelling
- Algorithmic Modelling

Data modelling supposes the existence of a stochastic model, conventional techniques such as linear regression, logistic regression, Cox regression are examples of data modelling. The data is used to estimate the parameters in the model. Algorithmic modelling does suppose some existing stochastic model; instead a black box approach is adopted. The independent variables X of the data model are considered as inputs to a black box which contains the unknown mechanism that generates the dependent variable(s) Y. The aim of data modelling is to find some function $f(x)$ that will predict y. The function $f(x)$ is an algorithm such as a neural network, or a support vector. Algorithmic modelling is not based on the principle of parsimony; Breiman argues that predictive accuracy demands a more complex prediction method, i.e. a more complex model; further Breiman (Breiman 2001) states that algorithmic models can provide better predictive accuracy than data models. There is with the algorithmic approach to modelling the problem of interpretability of the resulting model. If the model is so complex as to be beyond interpretation what good is this to a clinician? Breiman argues that it is still possible to acquire useful information the independent and the dependent variables. A distinction is made between information and interpretability, it might be that a simple model is easy to interpret but provides no

'real' information about the relationship between the independent and dependent variables, I can picture the model, but it is the wrong picture, or too simple a picture. I find Breiman's arguments extremely interesting.

In certain circumstances would there be a genuine benefit in using an Algorithmic approach in developing a prognostic model as opposed to the 'traditional' data modelling approach? Do Algorithmic models offer an advantage in so far as a reduced risk of over-fitting is concerned? Are Algorithmic models inherently less prone to over-fitting? Is it possible with an Algorithmic model to build into it a mathematical 'resistance' to over-fitting? Neural networks are certainly prone to over-fitting (Lawrence *et al.* 1997) as are support vector machines (Mierswa 2007). The present author would very much like to pursue an investigation of Algorithmic modelling techniques and compare them against comparable data modelling techniques.

9.3.0 MDL

An exciting approach to model selection which may overcome the problem of over-fitting is MDL (the Minimum Description Length). MDL has its origins in the theory of algorithmic complexity and information theory. In the MDL context a statistical model is considered as a description of the data, model selection is then based on the idea of choosing the smallest description. If a data set possesses regularity then it is possible to compress the data. By compress is meant the idea that the data can be described using less symbols or characters than would be needed to provide a literal description. The size of the description depends upon the detecting regularity within the data, the more regularity that the data exhibits the smaller the description, i.e. the smaller the model. The process of finding patterns or regularity within a data set is known as learning the data. Hansen and Yu (Hansen & Yu 2003) point out a major deficiency in model selection based on maximum likelihood, i.e. that the largest

model is the preferred choice. In chapter 5 of this work the AIC was discussed as a model selection tool, the AIC introduced a penalty term in order to correct the maximum likelihood model selection process. Hansen and Yu (Hansen & Yu 2003) state that the AIC performs well as a model selection tool if the underlying model is known to be of infinite dimension, but we do not generally have this information. MDL is proposed as a model selection method that is independent of the underlying model, and so is described as an adaptive method. The claim that MDL automatically protects against over-fitting (Rissanen 1978) can also be made for the AIC (due to the penalty term), the fact that MDL does not require the assumption of some underlying 'true' model is highly attractive feature. In the same way MDL may have benefits over the BIC, in the sense that the BIC performs well if the underlying model is of finite dimension, again for the BIC a 'true' underlying model is assumed. Data compression is a fundamental idea in MDL methods, there is a relationship between data compression and probability (this relationship can be expressed through Kraft's inequality (Kraft 1949)) this leads to the idea that MDL methods search for a model with good predictive power on new unseen data (Rissanen 1978). MDL is also related to cross validation (Rissanen 1978). MDL unlike the Algorithmic modelling discussed by Breiman (Breiman 2001) is based on Occam's razor, and so aims at a parsimonious model. The present author considers MDL as a potentially serious alternative to Harrell's approach. A comparison of models produced using MDL methods against those produced using Harrell's approach would be a most interesting project. The automatic protection against over-fitting afforded by MDL is of considerable benefit. With Harrell's approach the onus is to a greater extent on the researcher so far as taking steps to reduce the risk of over-fitting is concerned. I have formed the impression that MDL may represent a more cohesive approach than

Harrell's. The following material provides useful information MDL methods (Rissanen 1986), (Rissanen 1987), (Grunwald 2004) and (Hansen & Yu 2001)

9.4.0 Recommendations

9.4.1 Statistical Training And Accessible Literature

I believe that for Harrell's approach to be widely and routinely adopted the key issues of over-fitting and optimism need to be explained in a way that is intelligible to the non-technical expert at the point when they begin learning about statistics. Introductory courses on statistical modelling should cover the topics of over-fitting and optimism as a matter of routine and in tandem with modules on regression. It appears that the issues of over-fitting, optimism and model validation come back to haunt researchers some while after they have learnt what a Cox proportional Hazards model is all about. Harrell el al's modelling strategy as described in (Harrell *et al.* 1996) can be hard to follow and understand, a clearer exposition aimed at the non-statistician could be developed.

9.4.2 User friendly software

Software such as Hmisc, the RCS macro and the MFP macro can be rather daunting. I can imagine that even fairly computer literate researchers might find them awkward to use. Efforts to develop a more user friendly integrated modelling package that incorporates cubic splines, fractional polynomials, imputation and validation methods would be of considerable value. Harrell's software does indeed combine cubic splines

imputation and validation methods; however there are instances when the software proves to be awkward or limited.

9.4.3 Investigation of MDA Methods

An investigation of MDA methods applied to prognostic models would be in my opinion a useful piece of work. I intend to investigate further the theoretical and simulation studies relating to the AIC and BIC in conjunction with material on MDA methods. This will be done with a view to clarifying what advantages MDA may present as a model selection tool.

9.4.4 Frailty Models

Further investigation of frailty models is also an area that I intend to explore. The survival package in GNU R offers the facility to fit frailty models; model fit is reported via the likelihood ratio test. A form of the AIC for the frailty model as discussed in Chapter 5 has been proposed by Do Ha et al. (Do Ha *et al.* 2007); I would be interested attempting to implement this form of the AIC in software. Application of MDA methods to frailty models is of considerable interest to me.

Appendices

Appendix 1.0.0 SAS CODE

```
proc univariate data=card.prognostic;
    var mitral_r IVMD ESVI GRF QRS supsys supdia BSA HeartRate
a4cLVEjectionFraction Roche Age;
run;
data card.progex;
   set card.prognostic;
   if mitral_r ^= '.' and mitral_r < 11. then mitral_grp=1;
   if mitral_r ^= '.' and (mitral_r >= 11 and mitral_r < 22) then
mitral_grp=2;
   if mitral_r ^= '.' and (mitral_r >=22 and mitral_r < 34) then
mitral_grp=3;
   if mitral_r ^= '.' and mitral_r >= 34 then mitral_grp=4;
   trmit=treat*mitral_r;
   lmit=log(mitral_r);
   trlmit=treat*lmit;
   pmit=1/(sqrt(mitral_r));
   tpmit=treat*pmit;
   tanmit=tan(mitral_r);
   if IVMD ^= '.' and IVMD < 31 then IVMD_grp=1;
   if IVMD ^= '.' and (IVMD >= 31 and IVMD < 49) then IVMD_grp=2;
   if IVMD ^= '.' and (IVMD >=49 and IVMD < 67) then IVMD_grp=3;
   if IVMD ^= '.' and IVMD >= 67 then IVMD_grp=4;
   trivm=treat*IVMD;
   ShIVMD=IVMD+60;
   lShIVMD=log(ShIVMD);
   trlShIVMD=treat*lShIVMD;

   if ESVI ^= '.' and ESVI < 93 then ESVI_grp=1;
   if ESVI ^= '.' and (ESVI >= 93 and ESVI < 119) then ESVI_grp=2;
   if ESVI ^= '.' and (ESVI >=119 and ESVI < 149) then ESVI_grp=3;
   if ESVI ^= '.' and ESVI >= 149 then ESVI_grp=4;
   tresv=treat*ESVI;
   lesv=log(ESVI);
   trlesv=treat*lesv;
   if GRF ^= '.' and GRF < 46 then GRF_grp =1;
   if GRF ^= '.' and (GRF >= 46 and GRF < 60) then GRF_grp =2;
   if GRF ^= '.' and (GRF >= 60 and GRF < 73) then GRF_grp =3;
   if GRF ^= '.' and GRF >= 73 then GRF_grp=4;
```

```
   trgrf=treat*GRF;
   lgrf=log(GRF);
   trlgrf=treat*lgrf;
   if QRS ^='.' and QRS < 152 then QRS_grp =1;
   if QRS ^='.' and (QRS >= 152 and QRS < 160) then QRS_grp =2;
   if QRS ^='.' and (QRS >= 160 and QRS < 180) then QRS_grp =3;
   if QRS ^='.' and QRS >= 180 then QRS_grp=4;
   trqrs=treat*QRS;
   lqrs=log(QRS);
   trlqrs=treat*lqrs;
   if supsys ^='.' and supsys < 105 then supsys_grp =1;
   if supsys ^='.' and (supsys >= 105 and supsys < 117) then
supsys_grp =2;
   if supsys ^='.' and (supsys >= 117 and supsys < 130) then
supsys_grp =3;
   if supsys ^='.' and supsys >= 130 then supsys_grp=4;
   trsup=treat*supsys;
   lsup=log(supsys);
   trlsup=treat*lsup;
   if supdia ^='.' and supdia < 60 then supdia_grp =1;
   if supdia ^='.' and (supdia >= 60 and supdia < 70) then supdia_grp
=2;
   if supdia ^='.' and (supdia >= 70 and supdia < 80) then supdia_grp
=3;
   if supdia ^='.' and supdia >= 80 then supdia_grp=4;
   trdia=treat*supdia;
   ldia=log(supdia);
   trldia=treat*ldia;
   if BSA ^='.' and BSA < 1.73 then BSA_grp=1;
   if BSA ^='.' and (BSA >= 1.73 and BSA < 1.88) then BSA_grp=2;
   if BSA ^='.' and (BSA >= 1.88 and BSA < 2.01) then BSA_grp=3;
   if BSA ^='.' and BSA >=2.01 then BSA_grp=4;
   trbsa=treat*BSA;
   lbsa=log(BSA);
   trlbsa=treat*lbsa;
   if HeartRate ^='.' and HeartRate < 60 then HeartRate_grp=1;
   if HeartRate ^='.' and (HeartRate >= 60 and HeartRate < 69) then
HeartRate_grp=2;
   if HeartRate ^='.' and (HeartRate >= 69 and HeartRate < 78) then
HeartRate_grp=3;
   if HeartRate ^='.' and HeartRate >=78 then HeartRate_grp=4;
   trhea=treat*HeartRate;
   lhea=log(HeartRate);
   trlhea=treat*lhea;
   if a4cLVEjectionFraction ^= '.' and a4cLVEjectionFraction < 22
then EF_grp =1;
   if a4cLVEjectionFraction ^= '.' and (a4cLVEjectionFraction >= 22
and a4cLVEjectionFraction < 25) then EF_grp =2;
   if a4cLVEjectionFraction ^= '.' and (a4cLVEjectionFraction >= 25
and a4cLVEjectionFraction < 29) then EF_grp =3;
   if a4cLVEjectionFraction ^= '.' and a4cLVEjectionFraction >= 29
then EF_grp=4;
   tra4c=treat*a4cLVEjectionFraction;
   la4c=log(a4cLVEjectionFraction);
   trla4c=treat*la4c;
   if Roche ^='.' and Roche < 744 then Roche_grp =1;
   if Roche ^='.' and (Roche >= 744 and Roche < 1814) then Roche_grp
=2;
   if Roche ^='.' and (Roche >= 1814 and Roche < 4198) then Roche_grp
=3;
   if Roche ^='.' and Roche >= 4198 then Roche_grp=4;
```

```
   trroc=treat*Roche;
   lroc=log(Roche);
   trlroc=treat*lroc;
   if Age ^= '.' and Age < 59 then Age_grp =1;
   if Age ^= '.' and (Age >= 59 and Age < 66) then Age_grp =2;
   if Age ^= '.' and (Age >= 66 and Age < 72) then Age_grp =3;
   if Age ^= '.' and Age >= 72 then Age_grp=4;
   trage=treat*Age;
   lage=log(Age);
   trlage=treat*lage;
run;
quit;

proc phreg data=card.progex;
     title 'phreg treat treat*mitral_r futime';
     model futime*primary(0)=treat mitral_r trmit /RL;
run;
quit;

proc phreg data=card.progex;
     title 'phreg treat treat*log(mitral_r) futime';
       model futime*primary(0)=treat lmit trlmit /RL;
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(Mital_r Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\mitral.sas),
   TIME=futime, status=primary,
   COV1=mitral_r,WHAT1=0,KNOTS1=11 22 34 66,
   COV2=trmit,WHAT2=0,KNOTS2=11 22 34 66,
   COV3=treat
);

proc phreg data=card.progex;
     title 'phreg treat treat*IVMD futime';
     model futime*primary(0)=treat IVMD trivm /RL;
run;
quit;
proc phreg data=card.progex;
     title 'phreg treat treat*log(IVMD)futime';
       model futime*primary(0)=treat lShIVMD trlShIVMD /RL;
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(IVMD Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\IVMD.sas),
   TIME=futime, status=primary,
   COV1=IVMD,WHAT1=0,KNOTS1=31 49 67 115,
   COV2=trivm,WHAT2=0,KNOTS2=31 49 67 115,
   COV3=treat
);
proc phreg data=card.progex;
     title 'phreg treat treat*ESVI futime';
     model futime*primary(0)=treat ESVI tresv /RL;
run;
quit;
proc phreg data=card.progex;
```

```
     title 'phreg treat treat*log(ESVI) futime';
       model futime*primary(0)=treat lesv trlesv /RL;
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(ESVI Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\ESVI.sas),
   TIME=futime, status=primary,
   COV1=ESVI,WHAT1=0,KNOTS1=93 119 149 295,
   COV2=tresv,WHAT2=0,KNOTS2=93 119 149 295,
   COV3=treat
);
proc phreg data=card.progex;
     title 'phreg treat treat*GRF futime';
     model futime*primary(0)=treat GRF trgrf /RL;
run;
quit;
proc phreg data=card.progex;
     title 'phreg treat treat*log(GRF) futime';
       model futime*primary(0)=treat lgrf trlgrf /RL;
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(GRF Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\GRF.sas),
   TIME=futime, status=primary,
   COV1=GRF,WHAT1=0,KNOTS1=46 60 73 125,
   COV2=trgrf,WHAT2=0,KNOTS2=46 60 73 125,
   COV3=treat
);
proc phreg data=card.progex;
     title 'phreg treat treat*QRS futime';
     model futime*primary(0)=treat QRS trqrs /RL;
run;
quit;
proc phreg data=card.progex;
     title 'phreg treat treat*log(QRS) futime';
       model futime*primary(0)=treat lqrs trlqrs /RL;
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(QRS Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\QRS.sas),
   TIME=futime, status=primary,
   COV1=QRS,WHAT1=0,KNOTS1=152 160 180 218,
   COV2=trqrs,WHAT2=0,KNOTS2=152 160 180 218,
   COV3=treat
);
proc phreg data=card.progex;
     title 'phreg treat treat*supsys futime';
     model futime*primary(0)=treat supsys trsup /RL;
run;
quit;
proc phreg data=card.progex;
     title 'phreg treat treat*log(supsys) futime';
```

```
      model futime*primary(0)=treat lsup trlsup /RL;
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(Supsys Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\Supsys.sas),
   TIME=futime, status=primary,
   COV1=supsys,WHAT1=0,KNOTS1=105 117 130 165,
   COV2=trsup,WHAT2=0,KNOTS2=105 117 130 165,
   COV3=treat
);
proc phreg data=card.progex;
     title 'phreg treat treat*BSA futime';
     model futime*primary(0)=treat BSA trbsa /RL;
run;
quit;
proc phreg data=card.progex;
     title 'phreg treat treat*log(BSA) futime';
      model futime*primary(0)=treat lbsa trlbsa /RL;
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(BSA Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\BSA.sas),
   TIME=futime, status=primary,
   COV1=BSA,WHAT1=0,KNOTS1=1.73 1.88 2.01 2.38,
   COV2=trbsa,WHAT2=0,KNOTS2=1.73 1.88 2.01 2.38,
   COV3=treat
);
proc phreg data=card.progex;
     title 'phreg treat treat*HeartRate futime';
     model futime*primary(0)=treat HeartRate trhea /RL;
run;
quit;
proc phreg data=card.progex;
     title 'phreg treat treat*log(HeartRate) futime';
      model futime*primary(0)=treat lhea trlhea /RL;
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(HeartRate Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\HeartRate.sas),
   TIME=futime, status=primary,
   COV1=HeartRate,WHAT1=0,KNOTS1=60 69 78 105,
   COV2=trhea,WHAT2=0,KNOTS2=60 69 78 105,
   COV3=treat
);
proc phreg data=card.progex;
     title 'phreg treat treat*a4cLVEjectionFraction futime';
     model futime*primary(0)=treat a4cLVEjectionFraction tra4c /RL;
run;
quit;
proc phreg data=card.progex;
     title 'phreg treat treat*log(a4cLVEjectionFraction) futime';
      model futime*primary(0)=treat la4c trla4c /RL;
```

```
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(a4cLVEjectionFraction Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\a4c.sas),
   TIME=futime, status=primary,
   COV1=a4cLVEjectionFraction,WHAT1=0,KNOTS1=22 25 29 43,
   COV2=tra4c,WHAT2=0,KNOTS2=22 25 29 43,
   COV3=treat
);
proc phreg data=card.progex;
     title 'phreg treat treat*Roche futime';
     model futime*primary(0)=treat Roche trroc /RL;
run;
quit;
proc phreg data=card.progex;
     title 'phreg treat treat*log(Roche)futime';
       model futime*primary(0)=treat lroc trlroc /RL;
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(Roche Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\Roche.sas),
   TIME=futime, status=primary,
   COV1=Roche,WHAT1=0,KNOTS1=744 1814 4198 26132,
   COV2=trroc,WHAT2=0,KNOTS2=744 1814 4198 26132,
   COV3=treat
);
proc phreg data=card.progex;
     title 'phreg treat treat*Age futime';
     model futime*primary(0)=treat Age trage /RL;
run;
quit;
proc phreg data=card.progex;
     title 'phreg treat treat*log(Age) futime';
       model futime*primary(0)=treat lage trlage /RL;
run;
quit;
%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(Age Spline),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\Age.sas),
   TIME=futime, status=primary,
   COV1=Age,WHAT1=0,KNOTS1=59 66 72 84,
   COV2=trage,WHAT2=0,KNOTS2=59 66 72 84,
   COV3=treat
);
quit;


%INC 'C:\splines\rcs.mac';
%RCS(
   TITLE=%STR(Card Sync),
   DATA=progex, DIRDATA=%STR(C:\prog_card_dat),
   PROGRAM=%STR(C:\prog_card_dat\card_splines\All_sig.sas),
   TIME=futime, status=primary,
```

```
   COV1=supsys,WHAT1=0,KNOTS1=105 117 130 165,
   COV2=trsup,WHAT2=0,KNOTS2=105 117 130 165,
   COV3=BSA,WHAT3=0,KNOTS3=1.73 1.88 2.01 2.38,
   COV4=trbsa,WHAT4=0,KNOTS4=1.73 1.88 2.01 2.38,
   COV5=mitral_r,
   COV6=trmit,
   COV7=IVMD,
   COV8=trivm,
   COV9=lroc,
   COV10=ESVI,
   COV11=GRF,
   COV12=HeartRate,
   COV13=a4cLVEjectionFraction,
   COV14=Age,
   COV15=QRS,
   COV16=treat


);

Code for Final Model

proc phreg data=card.progex3;

       class Ischemic treat /desc;
       model futime*primary(0)= treat mitral_r IVMD ESVI GRF supsys
a4cLVEjectionFraction Roche Age Ischemic trsup trivm /RL
selection=forward slentry=0.5  details;

run;
proc phreg data=card.progex3;

       class Ischemic treat /desc;
       model futime*primary(0)= treat mitral_r Roche supsys IVMD
Ischemic trsup trivm  /RL details;
     baseline covariates=card.progex3 out=card.PrScore2
survival=S/nomean;
run;
```

References

Abraham W.T., Fisher W.G., Smith A.L., Delurgio D.B., Leon A.R., Loh E., Kocovic D.Z., Packer M., Clavell A.L., Hayes D.L., Ellestad M., Trupp R.J., Underwood J., Pickering F., Truex C., McAtee P., & Messenger J. 2002. MIRACLE Study Group. Cardiac resynchronization in chronic heart failure. **N Engl J Med** 346, 1845-1853.

Abu A & Lucas P.J.F. 2001. Prognostic Models in Medicine. **Method Inform Med** 40, 1-4.

Akaike H. 1974. A new look at the statistical model identification. **IEEE TransAutom Control** 19, 716-723.

Altman Douglas G. & Royston Patrick. 2000. What do we mean by validating a prognostic model? **Statistics in Medicine** 19, 453-473.

Altman Douglas G. & Royston Patrick. 2006. **The cost of dichotomising continuous variables**. BMJ 332[1080].

Altmayer Lawrence. 2009. Hot-Deck Imputation: A Simple DATA Step Approach.. U.S. Bureau of the Census, Washington. http://analytics.ncsu.edu/sesug/1999/075.pdf Last Accessed 20th April 2009.

Ambler Gareth, Brady Anthony R., & Royston Patrick. 2002. Simplifying a prognostic model: a simulation study based on clinical data. **Statistics in Medicine** 21, 3803-3822.

Auricchio A., Klein H., & Spinelli J. 1999. Pacing for heart failure: selection of patients, techniques and benefits. **Eur J Heart Fail** [1], 275-279.

Baker R.J. & Nelder J.A. 1978. **The GLIM SYSTEM RELEASE 3**, GENERALIZED LINEAR INTERACTIVE MODELLING MANUAL.

Beale E.M.L. 1970. Note on Procedures for Variable Selection in Multiple Regression. **Technometrics** 12[4], 909-914.

Benoît Minisini Website. 2009. **Gambas**. [2.12]. http://gambas.sourceforge.net/. Last accessed 20th April 2009.

Blundell Stephen J. & Blundell Katherine M. 2006. **Concepts in Thermal Physics**. Oxford University Press.

Boltzmann L. 1872. Weitere Studien uber das Warmegleichgewicht unter Gasmolekulen. **Sitzungsberichte der Akademie der Wissenschaften**,Wien II[66], 275-370.

Box George E.P. & Draper Norman R. 1987. **Empirical Model-Building and Response Surfaces.** 424. John Wiley & Sons Inc.

Bozdogan Hamparsum. 1987. Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. **Psychometrica** 52[3], 345-370.

Breiman Leo. 2001. Statistical Modeling: The Two Cultures. **Statistical Science** 16[3], 199-215.

Breslow, N. E. 1972. Discussion of Professor Cox's Paper, **Journal of the Royal Statistical Society.Series B (Methodological)**, 34, 216–217.

Bristow M.R., Saxon L.A., Boehmer J., Krueger S., Kass D.A., De Marco T., Carson P., , D. L., , D. D., White B.G., , D. D. W., & Feldman A.M. 2004. for the Comparison of Medical Therapy, Pacing, Defibrillation in Heart Failure (COMPANION) InvestigatorsCardiac-resynchronization therapy with or without an implanTable defibrillator in advanced chronic heart failure. **N Engl J Med** 350, 2140-2150.

Buck S.F. 1960. A Method of Estimation of Missing Values in Multivariate Data SuiTable for use with an Electronic Computer. **Journal of the Royal Statistical**

Society.Series B (Methodological) 22[2], 302-306.

Buckland S.T., Burnhan K.P., & Augustin N.H. 1997. Model Selection: An Integral Part of Inference. **Biometrics** 53[2], 603-618.

Burnham Kenneth P. & Anderson David R. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. **Sociological Methods Research** 33.

Cercignani Carlo. 2007. **Ludwig Boltzmann, The Man Who Trusted Atoms**. Oxford University Press.

Chakrabarti C.G. & Chakrabarty I. 2006. Boltzman-Shannon Entropy:Generalization and Application. http://arxiv.org/abs/quant-ph/0610177 Mod.Phys.Lett.B 20[23], 1471. Last accessed 20th April 2009.

Chao Celia, Studts Jamie L., Abell Troy, adley Terence, oetzer Lynne, ineen Sean, orenz Doug, oussefAgha Ahmed, & cMasters Kelly M. 2003. Adjuvant Chemotherapy for Breast Cancer: How Presentation of Recurrence Risk Influences Decision-Making. **Journal of Clinical Oncology** 21[23], 4299-4305.

Cleland J.G.F., Daubert J.C., Erdmann E., Freemantle N., Gras D., Kappenberger L., Klein W., & Tavazzi L. 2001. on behalf of the CARE-HF Study Steering Committee Investigators. The CARE-HF study (CArdiac Resynchronization in Heart Failure study): rationale, design and end-points. **Eur J Heart Fail** 3, 481-489.

Cleland J.G.F., Daubert J.C., Erdmann E., Freemantle N., Gras D., Kappenberger L., & Tavazzi L. 2005. Cardiac Resynchronization-Heart Failure (CARE-HF) Study Investigators. The effect of cardiac resynchronization on morbidity and mortality in heart failure. **N Engl J Med** 352, 1539-1549.

Cleland J.G.F., Gemmel I., Khand A., & Boddy A. 1999. Is the prognosis of heart failure improving? **Eur J Heart Fail** 1, 229-241.

Cox D.R. 1959. The Analysis of Exponentially Distiributed Life-times with Two Types of Failures. **Journal of the Royal Statistical Society.Series B (Methodological)**, 21, 411-421.

Cox D.R. 1964. Some Applications of Exponentially Distributed Life-times with Two Types of Failures. **Journal of the Royal Statistical Society.Series B (Methodological)**, 26, 103-110.

Cox D.R. 1972. Regression Models and Life Tables. **Journal of the Royal Statistical Society**, 34, 187-220.

Daubert J.C., Cazeau S., & Leclercq C. 1999. Do we have reasons to be enthusiastic about pacing to treated advanced heart failure? **Eur J Heart Fail** 1, 281-287.

Dempster A.P., Laird N.M., & Rubin D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal of the Royal Statistical Society.Series B (Methodological)** 39[1], 1-38.

Design Library Harrell Frank E. 2009a. **Overview of Design Library**. http://lib.stat.cmu.edu/S/Harrell/help/Design/html/Overview. Last accessed 20th April 2009.

Design Library Harrell Frank E. 11-4-2009b. **The Design Library**. [2.1-2]. http://biostat.mc.vanderbilt.edu/s/Design. Last accessed 20th April 2009.

Do Ha Il, Lee Youngjo, & Mackenzie Gilbert. 2007. Model selection for multi-component frailty models. **Statistics in Medicine** 26, 4790-4807.

Dobson Annette J. 2002. **An Introduction to Generalized Linear Models**. 2nd. Chapman Hall. Texts in Statistical Science.

Doust J.A., Pietrzak E., Dobson A., & Glasziou P. 2005. How well does B-type natriuretic peptide predict death and cardiac events in patients with heart failure: systematic review. **BMJ** 330, 625-635.

Efron B. 1979. Bootstrap Methods: Another Look At The Jacknife. **Anals of Statistics** 7, 1-26.

Efron B & Stein C. 1981. The Jackknife Estimate of Variance**. The Annals of Statistics** 9[3], 586-596.

Efron Bradley. 2000. The Bootstrap and Modern Statistics. **Journal of the American Statistical Association** 95[452], 1293-1296.

Efron Bradley & Gong Gail. 1983. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation**. The American Statistician** 37[1], 36-48.

Ellenbogen Kenneth A., Wood Mark A., & Klein Helmut U. 2005. Why Should We Care About CARE-HF? **Journal of the American College of Cardiology** 46[12], 2199-2203.

Euroscore Website. **The official website of the euroSCORE cardiac surgery** scoring system. http://www.euroscore.org/calc.html. Last accessed 20th April 2009.

Fernandez George. 2007. **Model Selection in PROC MIXED - A User-friendly SAS® Macro Application**. http://www2.sas.com/proceedings/forum2007/191-2007.pdf. Last Accessed 7th May 2009

Fisher R.A. 1932. Inverse Probability and the use of Likelihood. *Proceedings of the Cambridge Philosophical Society*, 28, 257-261

Fisher R.A. 1934a. Two New Properties of Mathematical Likelihood. *Proceedings of the Royal Society,A*, 144, 285-307

Fisher R.A. 1934b. Probability, Likelihood and Quantity of Information in the Logic of Uncertain Inference. *Proceedings of the Royal Society,A*, 146, 1-8

Fisher R.A. 1966. The Design of Experiments. 8th Edition. Oliver and Boyd

Freemantle N., Tharmanathan P., Calvert M.J., Abraham W.T., & Ghosh J. 2006. Cardiac resynchronisation for patients with heart failure due to left ventricular systolic dysfunction—a systematic review and meta-analysis. **Eur J Heart Fail** 8, 433-440.

Gaetan Carlo & Yao Jian-Feng. 2003. Multiple-Imputation Metropolis Version of the EM Algorithm. **Biometrika** 90[3], 643-654.

Gail M.H., Wieand S., & Piantadosi S. 1984. Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates. Biometrika 71[3], 431-444.

Ghio S., Constantin C., Klersy C., Serio A., Fontana A., Campana C., & Tavazzi L. 2004. Interventricular and intraventricular dyssynchrony are common in heart failure patients, regardless of QRS duration. **European Heart Journal** 25, 571-578.

Ginzburg Lev R. & Jensen Christopher X.J. 2004. Rules of thumb for judging Ecological Theories. **TRENDS in Ecology and Evolution** 19[3].

Grunwald Peter. 2004. **Tutorial Introduction to the Minimum Description Length Principle**. http://arxiv.org/PS_cache/math/pdf/0406/0406077.pdf Last Accessed 20th April 2009.

Hansen Mark H. & Yu Bin. 2001. Model Selection and the Principle of Minimum Description Length. **Journal of the American Statistical Association** 96[454], 746-774.

Hansen Mark H. & Yu Bin. 2003. Minimum Description Length Model Selection Criteria for Generalized Linear Models. Lecture Notes-Monograph Series, **Statistics and Science**: A Festschrift for Terry Speed. 40, 145-163.

Harrell Frank E., Lee Kerry L., & Mark Daniel B. 1996. Tutorial in Biostatistics Multivariate Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. **Statistics in Medicine** 15, 381-387.

Heinzl H. & Kaider A. 1997. Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. **Computer Methods and Programs in Biomedicine** 54, 201-208.

Heinzl H. & Kaider A. 2006. **Manual for the SAS-Macro RCS (Version 2. 0)**. Technical Report KB 1-96. Universitat Wien, Institut fur Medizinische Computerwissenschaften Abteilung fur Klinische Biometrie. http://www.meduniwien.ac.at/imc/biometrie/programme/Rcs.htm. Last Accessed 20th April 2009.

Heitjan Daniel F. & Little Roderick J.A. 1991. Multiple Imputation for the Fatal Accident Reporting System. **Applied Statistics** 40[1], 13-29.

Herndon James E. & Harrell Frank E. 1990. The Restricted Cubic Spline Hazard Model. **Commun.Statist.-Theory Meth** 19[2], 639-663.

Hmisc Library Harrell Frank E. 2009. **The Hmisc Library**. http://biostat.mc.vanderbilt.edu/s/Hmisc. Last Accessed 20th April 2009.

Horn K.P. 1983. The Curve of Least Energy. **ACM Transactions on Mathematical Software** 9[4], 441-460.

Horton Nicholas J. & Kleinman Ken P. 2009. Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. **The American Statistician** 61[1], 79.

Hosmer D.W. & Lemeshow S. 1992. **Applied Survival Analysis: Regression of Time to Event Data.** Wiley.

Hougaard Philip. 1984. Life Table Methods for Hetrogenous Populations; Distributions Describing the Hetrogeneity. **Biometrika** 71[1], 75-83.

Hougaard Philip. 1991. Modelling Heterogeneity in Survival Data. **Journal of Applied Probability** 28[3], 695-701.

Hurvich Clifford M. & Tsai Chih-Ling. 1995. Model Selection for Extended Quasi-Likelihood Models in Small Samples. **Biometrics** 51[3], 1077-1084.

ICH E9. 1999. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials (ICH E9), Expert Working Group. **Stat Med** 18, 1905-1942.

ICH HARMONISED TRIPARTITE GUIDELINE. 1998. **STATISTICAL PRINCIPLES FOR CLINICAL TRIALS** . http://www.ich.org/LOB/media/MEDIA485.pdf. Last Accessed 20th April 2009.

Justice Amy C., Covinsky Kenneth E., & Berlin Jesse A. 1999. Assessing the Generalizability of Prognostic Information. **Ann Intern Med** 130, 515-524.

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

Kendall M.G. 1938. A new measure of rank correlation. **Biometrika** 30, 81-93.

Kraft Leon Gordon. 1949. **A device for quantizing, grouping, and coding amplitude-modulated pulses**. Massachusetts Institute of Technology. Dept. of Electrical Engineering.

Kreyszig Erwin. 1993. **Advanced Engineering Mathematics**. 7th ed. John Wiley & Sons.

Kullback S. & Leibler R.A. 1951. On information and sufficiency. **Annals of Mathematical Statistics** 22, 79-86.

Lawrence Steve, Giles C.Lee, & Chung Tsoi Ah. 1997. Lessons in Neural Network Training: Overfitting May be Harder than Expected. **Proceedings of the Fourteenth National Conference on Artificial Intelligence** AAAI-97, 540-545.

Lee E.T. 1992. **Statistical Methods For Survival Data Analysis**. 2nd. John Wiley & Sons, Inc.

Lin, D., Wei, L. J., and Ying, Z. (1993), Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals. ***Biometrika***, 80, 557–572.

Little Roderick J.A. 1988a. A Test of Missing Completely at Random for Multivariate Data with Missing Values. **Journal of the American Statistical Association** 83[404], 1198-1202.

Little Roderick J.A. 1988b. Missing-Data Adjustments in Large Surveys. **Journal of Business & Economic Statistics** 6[3], 287-296.

London School of Hygiene and Tropical Medicine Website . 30-10-2008. **Missing Data**. http://www.lshtm.ac.uk/msu/missingdata/jargon_web/index.html .Last Accessed 20th April 2009.

Mantel N. 1970. Why stepdown procedures in variable selection. **Technometrics** 12, 621-625.

Manton Kenneth G., Stallard Eric, & Vaupel James W. 1986. Alternative Models for the Heterogeneity of Mortality Risks Among the Aged. **Journal of the American Statistical Association** 81[395], 635-644.

Marshall Andrea, **Contructing Prognostic Models In The Presence Of Missing Covariate Data**, Phd Thesis School of Mathematics ,University of Birmingham, , November 2007

Medtronic. 14-4-2009. **What is Cardiac Resynchronization Therapy?** http://www.medtronic.com/physician/hf/ Last Accessed 20th April 2009.

Meier-Hirmer Carolina, Ortseifen Carina, & Sauerbrei Willi. 2003. **Multivariable Fractional Polynomials in SAS**. An Algorithm for Determining the Transformations of Continuous Covariates and Selection of Covariates. Institute of Medical Biometry, Freiburg Germany http://www.imbi.uni-freiburg.de/biom/mfp/. Last Accessed 20th April 2009.

MFP. 2009. **MFP**. http://www.imbi.uni-freiburg.de/biom/mfp/index.html. Last Accessed 20th April 2009.

Mierswa Ingo. 2007. Controlling Overfitting with Multi-Objective Support Vector Machines. ***GECCO'07.*** **ACM** 978-1-59593-697-4/07/0007.

Miller Rupert G. 1974. The Jackknife-A Review. **Biometrika** 61[1], 1-15.

Montori Victor M. & Guyatt Gordon H. 2001. Intention-to-treat principle. **CMAJ** 165[10].

Moons Karel G M., Royston Patrick, Vergouwe Yvonne, Grobbee Diederick E, & Altman Douglas G. 2009. Prognosis and prognostic research: what, why, and how? **BMJ** 338.

Nariaki Sugiura. 1978. Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections. **Communications in Statistics, Theory and Methods** A7, 13-26.

Nelder J.A. & Wedderburn R.W.M. 2009. Generalized Linear Models. **Journal of the Royal Statistical Society, Series A** 135, 370-384.

NICE. 2009. **NICE**. http://www.nice.org.uk/. Last Accessed 20th April 2009.

Nielsen Søren Feodor. 2003. Proper and Improper Multiple Imputation. **International Statistical Review / Revue Internationale de Statistique** 71[3], 593-607.

Pagano D., Freemantle N., Bridgewater B., Howell N., Ray D., Jackson M., Fabri B.M., Au J., Keenan D., Kirkup B., & Keogh B.E. 2009. Social deprivation and prognostic benefits of cardiac surgery: observational study of 44 902 patients from five hospitals over 10 years. **BMJ** 338:b902.

Pawitan Yudi. 2001. **In All Likelihood: Statistical Modelling and Inference Using Likelihood.** Oxford University Press.

Peduzzi P., Concato J., Kemper E., Holford T.R., & Feinstein A.R. 1996. A simulation study of the number of events per variable in logistic regression analysis. **Journal of Clinical Epidemiology** 49[1503], 1510.

Perperoglou Aris, Keramopoullos Antonis, & van Houwelingen Hans C. 2007. Approaches in Modelling long-term survival: An Application to breast cancer. **Statistics in Medicine** 26, 2666-2685.

Picard Richard R. & Berk Kenneth N. 1990. Data Splitting. **The American Statistician** 44[2], 140-147.

Pitzalis M.V., Iacoviello M., Romito R., Guida P., De Tommasi E., Luzzi G., Anaclerio M., Forleo C., & Rizzon P. 2005. Ventricular asynchrony predicts a better outcome in patients with chronic heart failure receiving cardiac resynchronization therapy. **J Am Coll Cardiol** 45, 65-69.

Poirier Dale J. 1979. Piecewise Regression Using Cubic Spline. **Journal of the American Statistical Association** 68[343], 515-524.

R Foundation for Statistical Computing, V. A. 2009. **R: A language and environment for statistical computing**. http://www.R-project.org Last Accessed 20th April 2009.

R survival package Terry Therneau. 2009. **R survival package**. http://www.stats.bris.ac.uk/R/web/packages/survival/index.html Last Accessed 20th April 2009.

Richardson Matthew, Freemantle Nick, Calvert Melanie J., Cleland John G.F., & Tavazzi Luigi. 2007. Predictors and treatment response with cardiac resynchronization therapy in patients with heart failure characterized by dyssynchrony: a pre-defined analysis from the CARE-HF trial. **European Heart Journal** 28, 1827-1834.

Rissanen J. 1978. Modeling by shortest data description. **Automatica** 14, 465-471.

Rissanen Jorma. 1986. Stochastic Complexity and Modeling. **The Annals of Statistics** 14[3], 1080-1100.

Rissanen Jorma. 1987. Stochastic Complexity**. Journal of the Royal Statistical Society.Series B (Methodological)** 49[3], 223-239.

Royall Richard M. 1991. Ethics and Statistics in Randomized Clinical Trials. **Statistical Science** 6[1], 52-62.

Royston Patrick & Altman Douglas G. 1994. Regression Using Fractional Polynomials of Continuous Covariates. **Applied Statistics** 43[3], 429-467.

Royston Patrick, Ambler Gareth, & Sauerbrei Willi. 1999. The use of fractional polynomials to model continous risk variables in epidemiology. **International Journal of Epidemiology** 28, 964-974.

Royston Patrick, Moons Karel G M., & Altman Douglas G. 2009. Prognosis and prognostic research: Developing a prognostic model. **BMJ** 338.

Royston Patrick & Sauerbrei Willi. 2004. A new approach to modelling interactions between treatment and continous covariates in clinical trials by using fractional polynomials. **Statistics in Medicine** 23, 2509-2525.

Royston Patrick. 2000. Choice of scale for cubic smoothing spline models in medical applications. **Statistics in Medicine** 19, 1191-1205.

Rubin Donald B. 1976. Inference and Missing Data. **Biometrika** 63[3], 581-592.

Rubin Donald B. 1996. Multiple Imputation After 18+ Years**. Journal of the American Statistical Association** 91[434], 473-489.

Rubin Donald B. 2003. Discussion on Multiple Imputation. **International Statistical Review / Revue Internationale de Statistique** 71[3], 619-625.

Runge Carl. 1901. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. **Zeitschrift für Mathematik und Physik** 46, 224-243.

SAS Institute. 2009. **The MI Procedure.** http://support.sas.com/rnd/app/papers/miv802.pdf Last Accessed 20th April 2009.

SAS Proc MI. 2009. **Proc MI**. http://www.sas.com/rnd/app/papers/miv802.pdf Last Accessed 20th April 2009.

SAS Proc Mianlyze. 2009. **Proc Mianlyze**. http://www.sas.com/rnd/app/papers/mianalyzev802.pdf Last Accessed 20th April 2009.

Schafer Joseph L. & Olsen Maren K. 1998. Multiple Imputation for Multivariate Missing-Data: A Data Analyst's Perspective. **Multivariate Behavioral Research** 35[4], 545-571.

Schoenberg I.J. 1966. On Hermite-Birkhoff interpolation**. Math.Anal.Appl** 10, 538-543.

Schwarz Gideon. 1978. Estimating the Dimension of a Model**. The Annals of Statistics** 6[2], 461-464.

Sedgwick James E.C. 2001. Absolute, attribuTable, and relative risk in the management of coronary heart disease**. Heart** 85, 491-492.

Shannon C.E. 1948. A Mathematical Theory of Communication. **The Bell System Technical Journal 27**, 379-423.

Sharma Subhash. 1995. **Applied Mulitvariate Techniques**. Wiley.

Smith Patricia L. 1979. Splines As a Useful and Convenient Statistical Tool. **The American Statistician** 33[2].

Somers Robert H. 1962. A New Asymmetric Measure of Association for Ordinal Variables. **American Sociological Review** 27[6], 799-811.

Stocken D.D., Hassan A.B., Altman D.G., Billingham L.J., Bramhall S.R., Johnson P.J., & Feemantle N. 2008. Modelling prognostic factors in advanced pancreatic cancer. **British Journal of Cancer** 99, 883-893.

Stone C.J. & Koo C.Y. 1986. Additive Splines in Statistics. Proceedings of the Statistical Computing Section of The American Statistical Association , Washington DC, **American Statistical Association**. 45-48.

Stone M. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. **Journal of the Royal Statistical Society.Series B (Methodological)** 36[2], 111-147.

Talwar S., Squire B., Davies J.E., Barnett D.B., & Ng L.L. 1999. Plasma N-terminal probrain natriuretic peptide and the ECG in the assessment of left ventricular systolic dysfunction in a high risk population. **European Heart Journal** 20, 1737-1777.

Therneau T.M. & Grambsch P.M. 1990. Martingale-based residuals for survival models. **Biometrika** 77, 147-160.

Ulrich Rich. 1997. **Stepwise comments - Harrell, Bernstein, Conroy**. http://www.pitt.edu/~wpilib/statfaq/regrfaq.html. Last Accessed 6th May 2009

Vaida Florin & Blanchard Suzette. 2005. Conditional Akaike information for mixed-effects models. **Biometrika** 92[2], 351-370.

van Houwelingen Hans C. 2000. Validation, calibration, revision and combination of prognostic survival models. **Statistics in Medicine** 19, 3401-3415.

Van Houwelingen J.C & le Cessie S. 1990. Predictive Value of Statistical Models. **Statistics in Medicine** 8, 1303-1325.

Vaupel James W., Manton Kenneth G., & Stallard Eric. 1979. The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. **Demography** 16[3], 439-454.

Verweij P.J.M., van Houwelingen H.C., & Stijnen T. 1998. A goodness of fit test for Cox's Proportional Hazards Model based on martingale residuals. **Biometrics** 54, 1517-1526.

Wahba G. & Wold S. 1975. Periodic splines for spectral density estimation: the use of cross validation for
determining the degree of smoothing. **Communications in Statistics - Theory and Methods** 4[2], 125-141.

Wegman Edward J. & Wright Ian W. 1983. Splines in Statistics. **Journal of the American Statistical Association** 78[382], 351-365.

Wienke Andreas. 2003. **Frailty Models**. http://www.demografie.eu/MPIDR WORKING PAPER WP 2003-032. Max Planck Institute for Demographic Research. Last Accessed 20th April 2009.

Wold Svante. 1974. Spline Functions in Data Analysis. **Technometrics** 16[1], 1-11.

Wyatt Jeremy C. & Altman Douglas G. 1995. Commentary: Prognostic models: clinically useful or quickly forgotten? **BMJ** 311, 1539-1541.

Xiao H.B., Brecker S.J.D., & Gibson D.G. 1993. Differing effects of right ventricular pacing and left bundle branch block on left ventricular function. **Br Heart J** 69, 166-173.

Zhang Paul. 2003. Multiple Imputation: Theory and Method. **International Statistical Review / Revue Internationale de Statistique** 71[3], 581-592.