# ESTIMATION OF NON-LINEAR AND TIME-VARYING EFFECTS IN SURVIVAL DATA USING SMOOTHING SPLINES

Author

ALBERT OTIENO ORWA

MASTER OF SCIENCE IN MATHEMATICS (STATISTICS OPTION)

PAN AFRICAN UNIVERSITY
INSTITUTE FOR BASIC SCIENCES TECHNOLOGY AND INNOVATION

September 2014

# Estimation of Non-Linear and Time-Varying Effects in Survival Data using Smoothing Splines

Albert Otieno Orwa

MS300-0007/2012

A thesis submitted to Pan African University, Institute for Basic Sciences Technology and Innovation in partial fulfillment of the requirement for the degree of

M.Sc. of Science in Mathematics (Statistics Option)

September 2014

# DECLARATION

This thesis is my original work and has not been submitted to any other university for examination.

Signature:..... Date:.....

Albert Otieno Orwa

This thesis report has been submitted for examination with our approval as University supervisors.

Signature:..... Date:.....

Dr. George Otieno Orwa

Department of Statistics & Actuarial Science

Jomo Kenyatta University

Signature:..... Date:.....

Prof. Romanus. O. Odhiambo

Department of Statistics & Actuarial Science

Jomo Kenyatta University

# ACKNOWLEDGEMENTS

I am most grateful to my tutors Dr. George Orwa and Professor Romanus Odhiambo for their invaluable guidance, support, stimulating ideas and the supervision both during my studies and in the conduct of the project. I would like to acknowledge all my course mates, friends and family for their encouragement and support through the year. Last but not least; I would not forget to thank the PAUSTI, for funding my studies and to the Almighty God for the gift of life and strength.

Thank you, Lord, for always being there for me.

This thesis is just a beginning of my journey.

# DEDICATION

I dedicate my dissertation work to my family. I also dedicate it to my many friends and church family who have supported me throughout the process. I will always appreciate all that they did towards accomplishing my goals.

I dedicate this work and give special thanks to my best friend Belinder Aoko and my wonderful daughter Hadassah for being there for me throughout the entire Masters program. Both of you have been my best cheer leaders.

# ABSTRACT

The major interests of survival analysis are either to compare the failure time distribution function or to assess the effects of covariate on survival via appropriate hazards regression models. Cox's proportional hazards model (Cox, 1972) is the most widely used framework, the model assumes that the effect on the hazard function of a particular factor of interest remains unchanged throughout the observation period (Proportionality assumption). For a continuous prognostic factor the model further assumes linear effect on the log hazard function (Linearity assumption). Assumptions that many authors have found to be questionable when violated since they may result to biased results and conclusions and as such non-linear risk functions have been suggested as the suitable models. In this paper, we propose a flexible method that models dynamic effects in survival data within the Cox regression framework. The method is based on penalized splines. The model offers the chance to easily verify the presence of PH and time-variation. We provide a detailed analysis and derivation of the penalized splines in the context of survival data.

**Key Words:** Non-linear, Penalized splines, Proportional Hazard, Survival analysis

# Table of Contents

# LIST OF TABLES

# LIST OF FIGURES

## LIST OF ABBREVIATIONS

ALHM : Aalen Linear Hazards Model

AMSE : Average Mean Squared Error

CPHM : Cox Proportional Hazard Model

DCR : Department of Civil Registration

FP : Fractional Polynomial

KM : Kaplan Meier

PH : Proportional Hazard

RCS : Restricted Cubic Splines

# CHAPTER 1
# INTRODUCTION

## 1.1 Background of the study

Survival analysis encompasses a wide variety of methods aimed at analyzing the timing of events. Many researchers are able to model and assess why certain subjects are exposed to a higher risk of experiencing an event of interest such as death, development of an adverse reaction or relapse of a given disease (e.g. Cancer).

Cox proportional hazard model is the most popular regression model used for the analysis of survival data. The model allows testing for the differences in survival times of two or more groups of interest and compares the cumulative probability of the events, while adjusting other influential covariates. It is a semi-parametric model that makes fewer assumptions than a typical parametric method. One of the assumptions of the Cox model is the linearity of the covariates variables on the log hazard function. The non-flexibility of these methods subjects the model to biasness. For instance, they assume independence of covariates that affect the hazard rate. They also assume that the model is linear yet findings have indicated that some prognostic factors (for example, body mass index) have non-linear effect on breast cancer survival and/or prognosis (Gray, 1994). Based on this, cox proportional model poses a problem in analyzing time-to-event data;

i) It is complex to relate the variables to the outcome.
ii) The variables interact with each other.
iii) It is not possible to apply the assumption of proportionality of the hazards to the data.

This could possibly lead to biased risk estimates and as such distorting the findings. (Hastie, T & Tibshirani, R, 1990) have shown that a better choice is to use smoothing splines, where knot selection is automatic based on a mean squared criterion. With smoothing splines, the user only need to select the level of smoothness, which is done by selecting the degrees of freedom for each spline fit.

## 1.2 Justification of the study

With a prevalence rate of 33.5 per cent, breast cancer has of late been described as one of the leading killer disease among Kenyan women (DCR report 2012). The use of smoothing methods will yield parsimonious models that will select significant variables hence revealing nonlinearities in the effects of predictors. The approach will be compared to standard methods by simulations and an example. It is imperative that accurate results that will help policy makers and other players in the field are obtained. The use of smoothing methods will help identify non-linear and time varying effects and the proposed model will help in coming up with unbiased results that are not subject to distortion.

## 1.3 Statement of the problem

Cox proportional hazard model is the most widely used framework in survival analysis where the main focus of analysis lies in modeling the impact of various prognostic factors and therapy on the time to occurrence of a given event/outcome for example death or relapse of a disease.

Although it is a well-designed and validated, this model do not consider time-varying effects of their covariates and as such assumes proportionality of the hazards and also that the risk factors act multiplicatively on the baseline hazard risk function. These assumptions, however, may not be proper in some applications and there is therefore the need for alternative models.

Risk factors may also have additive effects instead of multiplicative effects in the baseline hazard function. Another typical deviation from the proportional hazards Cox model is when the effects of some covariates change with time. For instance, some risk factors may impose a strong effect right after being recorded, but gradually lose predictive power (e.g. a treatment effect that is weakened with time). Models flexible enough to deal with covariates in which their effects are time-varying are therefore of great interest in these situations.

## 1.4 Objectives

The main objective of this research will be to propose a non-linear model for estimating time varying effects in survival data based on smoothing splines. However, this objective will further be supplemented with two specific objectives listed below;

1. Explore and identify non-linear and time varying effects in survival data
2. Propose a survival model for estimating the non-linear and time varying effects.
3. Study the properties of the proposed non-linear model (Smoothing splines)
4. Apply the empirical survival data to the model.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

Time-varying effects (TVEs) of prognostic factors have been detected in a variety of medical fields. Gore et.al (1984) presents a classical example discussing this issue for several covariates that relate to breast cancer. In the same disease, the effects of oestrogen receptor and tumour size have been reported to change over time (Hilsenbeck, S. G., et.al, 1998). Other examples have been established to include the effects of prothrombin time in primary biliary cirrhosis (Abrahamowicz, M., , T. MacKenzie,, & J. M. Esdaile, 1996), the Karnofsky performance status in ovarian cancer studies (Verweij & Houwelingen, 1995) and diabetes on mortality after coronary artery bypass graft surgery (Gao C, , Yang M,, Wu Y, , & et al, 2006).

Non-linearity is modeled with time transformations known as fractional polynomials (FPs) having power terms that can be negative values and fractions with conventional polynomials (CPs) as a special case (Long J, & Ryoo J, 2010). They (Long J, & Ryoo J, 2010), further showed in their results that the FPs had better or rather equal fit than the higher-order CPs and had prediction curves with as favorable or more favorable characteristics, such as less extreme behavior at the edges of the observed time intervals.

## 2.2 Cox proportional hazard (CPH) model

The semi-parametric proportional hazards model of (Cox, 1972) has become the standard for the analysis of survival time data in cancer studies and in many other application areas in medicine. In most studies, proportional hazards (PH) are assumed for covariate effects, implying that the effect on the hazard function of therapy and of each potential prognostic factor measured at the beginning of a study is unchanged throughout the whole observation period. However, with long-term follow-up this assumption may be questionable. For example, in a study of breast cancer patients, (Hilsenbeck, S. G., et.al, 1998), demonstrated that several factors violate the PH

assumption. Tumour size was established to have a strongly influence on the short term prognosis, but the effect diminished over time and was less relevant for prognosis after a patient had remained disease-free for a longer period. The effect of oestrogen receptor (ER) status and S-phase fraction also varied in time. Time-varying effects of tumour size and ER status have also been reported in breast cancer by others, for example (Coradini, D., et.al, 2000).

Well-known methods for checking the PH assumption have been available for some time. (Hess, 1994; Ng'Andu, 1997; Berger, U., , J. Schäfer, , & K. Ulm, 2003), provide an overview and some comparisons of the test statistics properties. Nevertheless, there is no agreement about which methods to use. Checking the PH assumption is often not even mentioned in papers describing applications of the Cox model.

By 1972, Cox had suggested relaxing the PH assumption by including an interaction between a covariate and a pre-specified parametric function of time. Typically, linear or logarithmic functions have been used. Since then, several other methods have been proposed for incorporating such a time dependent function for a covariate (e.g. for a prognostic factor). Examples include a step-function model based on cut points on the time axis, the use of smoothing splines (Hastie, T. & Tibshirani, R., 1993), penalized regression splines (Gray R. J., 1992), regression splines (Hess, 1994; Heinzl, H. & A. Kaider, 1997) and fractional polynomials (Berger, U., , J. Schäfer, , & K. Ulm, 2003). Other authors have investigated time-varying effects with even more flexible approaches by estimating local time-varying coefficients (Verweij & Houwelingen, 1995; Cai, Z. & Y. Sun, 2003; Martinussen, T., , T. H. Scheike, , & I. M. Skovgaar, 2002).

## 2.3 Parametric functions of time

### 2.3.1 Time-dependent covariate method

Cox in his original paper (1972) proposed extension of the Cox proportional Hazard model. He introduced time-dependent components that utilized pre-defined time functions in case of non-PH. To check the proportionality assumption, it is thus

proposed to fit an extended cox model containing time-dependent variables which is defined using some time function.

This involves inclusion of a time-dependent covariate $X_i f_i(t)$ that represents an interaction between the time parametric function. This corresponds to the inclusion ($f_i(t)$) and the predictor. The model is thus modified as;

$$\lambda(t|X) = \lambda_0(t)exp\left(\sum_{i=1}^{q} X_i f_i(t)\beta_i\right) = \lambda_0(t)exp\left(\sum_{i=1}^{q} X_i \beta_i(t)\right) \qquad (1)$$

Where, $\beta_i(t) = \gamma_{i0} + \gamma_{i1} f_i(t)$, represents the time-varying effects. The above model provides assessment of the PH assumption by testing the null hypothesis $\gamma_{i0} = 0$ by computing the likelihood ratio test statistic. The limitation of the model lies in making inferences of the test results however, the implementation of the model is very easy using the standard statistical software such as R and Stata.

### 2.3.2 Linear correlation test

A simple test based on Schoenfeld's partial residuals of the model for assessing the Cox's PH assumption was developed by Harrel. The test is based on the Pearson correlation between rank order of the failure time and the partial residuals.

The residuals neither depend on time nor do they involve estimation of the baseline hazard function. In the presence of tied failure times, the residual is taken as the total component of the first derivatives of the log-likelihood function based on the regression parameter. To check whether PH assumption holds, we test the null hypothesis ($H_0: \rho = 0$) using the formula $z = \rho\sqrt{(n_u - 2)/(1 - \rho^2)}$, where $n_u$ represents the total number of uncensored observations and $\rho$ represents the correlation between failure time order and residuals. The test statistic tends to be positive if the ratio of the hazards for high values of the covariate increases over time, and it tends to be negative if this hazard ratio decreases over time.

## 2.4 Smoothing splines

Smoothing spline methods are becoming popular modeling tools in many survival data contexts. They make it possible to handle complex non-linear relationships that are otherwise considered difficult to be estimated by the conventional parametric models. Splines are known for their ability to render good approximations to smooth functions (Boor, 1978; Schumaker, 1981) and their application in nonparametric smoothing is broad (Stone, 1997). Spline application in the current context is of particular convenient. After choosing a spline basis, the spline is approximated by unknown. The regression coefficients and the coefficients of spline are then simultaneously estimated by maximizing the partial likelihood.

The spline estimate $\hat{\mu}$ related to the function $\mu$ is defined as the minimizer (over the class of twice differentiable functions) of;

$$\sum_{i=1}^{n}(Y_i - \hat{\mu}(x_i))^2 + \lambda \int_{x_1}^{x_n} \widehat{\mu}''(x)^2 dx \qquad (2)$$

The number of different approaches to spline smoothing is quite wide, ranging from smoothing spline techniques (Hastie, T & Tibshirani, R, 1990; Wahba, 1990; Green, D. J. & Silverman, B. W., 1994), where a knot is placed at each observation, to regression splines with adaptive knot selection (Friedman, 1991). Eilers, P. H and Marx, B. D (1996) proposed the application and use of the penalized splines, a different approach which can be seen as a compromise between smoothing and regression splines. There, the number of knots defining the spline function is larger than that justified by the data, but smaller than the number of observations. According to Bulcholz. A (2010), splines represents a large group of approaches used in modeling time-varying effects. Splines are flexible non-parametric tools that are used to identify functional relationships and produce smooth visible curves. The construction of splines is based on joining polynomial pieces at certain values called knots. The fitted curve is influenced by the choice of position and the number of the knots. Very many knots may result to overfitting of the data, while very few knots may result in an underfitting (Buchholz, 2010). Solutions to this problem tend to two directions. Either relatively few knots are used or a relatively large number of knots is

combined with a smoothness penalty, resulting in penalized splines (Eilers, P. H. C. & Marx, B. D., 1996).

For example, (Hess, 1994) and (Heinzl, H. & A. Kaider, 1997) used (unpenalized) natural cubic splines with 3 to 5 knots. Their proposals include a formal test of the PH assumption by testing on the spline coefficients being equal to zero. (Abrahamowicz, M., , T. MacKenzie,, & J. M. Esdaile, 1996) prefer quadratic B-splines with no more than two knots. (Hastie, T. & Tibshirani, R., 1993) proposed a penalized partial likelihood approach based on natural cubic splines, with knots at unique event times and second order penalty based on the squared second derivative of the time-varying effects. The values of the smoothing parameters are selected by specifying the degrees of freedom for the smooth, i.e. the effective number of parameters. (Gray R. J., 1992) uses a similar method to determine the smoothing parameters, but bases the estimation of time-varying effects on B-splines of degree two and zero (i.e. piecewise constant effects) with a first order integral and first order difference penalty, respectively. The number of knots is limited to ten. (Brown, D., , G. Kauermann, , & I. Ford, 2007) proposed a mixed model approach, which assumes that some effects are random. They use linear B-splines or, equivalently, truncated polynomials penalized by a difference matrix or identity matrix, respectively, to approximate the time-varying effects. The smoothing parameters relate to the variance components in the mixed model framework and are estimated by method of cross validation. The estimation procedure cycles between estimating the regression coefficients for given smoothing parameters and vice versa, checking the AIC at each and every iteration. The procedure stops if the AIC can no longer be improved.

### 2.4.1 Penalized smoothing splines

Penalized spline (P-spline) smoothing is discussed for hazard regression of multivariable survival data. Non-proportional hazard functions are fitted in a numerically handy manner by employing Poisson regression which results from numerical integration of the cumulative hazard function. Multivariate smoothing parameters are selected by utilizing the connection between P-spline smoothing and generalized linear mixed models. A hybrid routine is suggested which combines the

mixed model idea with a classical Akaike information criteria (AIC). The model is evaluated with simulations and applied to data on the success and failure of newly founded companies.

### 2.4.2 Fractional Polynomial (FP)

Fractional polynomial (FP) regression models are the link between polynomial and nonlinear models. The aim in using FP functions in regression is to keep the advantages of conventional polynomials, while eliminating the disadvantages. FP functions are identical to conventional polynomials in that they include powers of X, however they don't allow non-integer and negative powers (Royston & Altman, 1994). FP models usually give a better fit than conventional polynomials of the same degree, and even than those of higher degree. FP functions can be used with any generalized linear model and with Cox proportional hazards regression models for survival data.

The FP of degree $m$ is the function

$$\emptyset_m(X;\xi,P) = \xi_0 + \sum_{j=1}^{m} \xi_j X^{(p_j)} \qquad (3)$$

where $m$ is a positive integer, $p = (p_1, p_2, \dots, p_m)$ is a real-valued vector of powers $p_1 < \cdots < p_m$ and $\xi = (\xi_0, \xi_1, \dots, \xi_m)$ are real-valued coefficients.

### 2.4.3 Restricted Cubic Spline (RCS)

Cubic splines are generally defined as piecewise-polynomial line segments whose function values and first and second derivatives agree at the boundaries where they join. The boundaries of these segments are called knots, and the fitted curve is continuous and smooth at the knot boundaries. To avoid instability of the fitted curve at the extremes of the covariate, a common strategy is to constrain the curve to be a straight line before the first knot or after the last knot. In this study, multivariate analysis of Cox PH model will be fitted on all variables to determine the effective factors on survival of the patients with breast cancer. Due to the suitability of spline models for continuous predictor variables, to compare the Cox PH model with P-

spline, fractional polynomial and restricted cubic spline in Cox model from identified continuous effective variables in multivariate Cox PH model will be used.

# CHAPTER 3

## METHODOLOGY

## 3.1 Review of regression models

### 3.1.1 Regression model

A regression model is basically of the form;

$$Y = f(X_1, X_2, \ldots, X_n) + \varepsilon \quad (4)$$

Where, Y is the response (dependent) variable $X_1, X_2, \ldots, X_n$ are the predictor (independent) variables, $\varepsilon$ is the error or simply the difference between the model and the actual values. The regression model aims at minimizing the error ($\varepsilon$) for all the values of $Y$ without introducing extraneous and arbitrary random variables. For a single predictor variable (univariate variable) we have;

$Y = f(X) + \varepsilon$ for some function $f$.

### 3.1.2 Simple linear regression

Simple linear regression or ordinary least squares (OLS) fits a straight line to the dataset of interest. It is given as;

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (5)$$

Where $\varepsilon$ is the error term (accounting for difference between the predicted and observed $Y$ values). We make assumptions that the error has a mean zero and a constant variance $\sigma^2$, and is identically independently distributed (iid). By this we mean that each error term is centered about the line of best fit (with mean zero) and that there is a constant amount of deviation of the error terms from the line of best fit

(with constant variance).To find the line of best fit through the scatter plot of $(X, Y)$ values, we actually aim at minimizing the error term for all the values of $y$. We then modify our equation as $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and the fitted or predicted value becomes, $\hat{y}_i = \beta_0 + \beta_1 x_i$ which simply implies that; $y_i = \hat{y}_i + \varepsilon_i$

We rewrite the model by solving the error term as;

$$\varepsilon_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i \tag{6}$$

Selecting the values of $\beta_0$ and $\beta_1$ that minimizes the total error as much as possible. We have;

$$Min \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{7}$$

Taking the partial derivatives and set them to zero

$$\frac{\partial}{\partial \beta_0} = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) \tag{8}$$

$$\frac{\partial}{\partial \beta_1} = \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i) \tag{9}$$

The above equations yield the following two normal equations;

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \tag{10}$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} {x_i}^2 \tag{11}$$

These are two equations in two unknowns. We can thus solve for $\beta_0$ and $\beta_1$ yielding;

$$\beta_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i\right) = \bar{y} - \beta_1 \bar{x} \tag{12}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} y_i - n\beta_0}{\sum_{i=1}^{n} x_i} \tag{13}$$

## 3.2 Derivation of spline regression model

Spline regression is a regression model with piecewise continuous polynomial function. We intend to derive penalized spline. Considering a simple linear model (5), and applying the concept of algebra we have; $\hat{y} = X\beta$ which can be rewritten in matrix form as follows;

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \beta = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \tag{14}$$

with $\hat{\beta} = (X^T X)^{-1} X^T Y$. Clearly $\hat{y}$ is a unique linear combination of the x-values and 1, the basis is thus x and 1.

### 3.2.1 Penalized Splines

Using penalization criteria we choose Q such that;

$$\sum_{i=1}^{k} b_i^2 < Q \tag{15}$$

The above equation represents a minimization criterion since it reduces the overall effect of individual piecewise functions and avoids over-fitting the data. We can formally state the minimization criterion as minimizing the equation given below;

$|y - X\beta|^2$ subject to $\beta^T D\beta \leq Q$, where;

$$D = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 0_{2\times 2} & 0_{2\times k} \\ 0_{k\times 2} & I_{k\times k} \end{bmatrix} \tag{16}$$

Applying Lagrange Multiplier results an equation which is equivalent to minimizing;

$$|y - X\beta|^2 + \lambda^2 \beta^T D\beta \tag{17}$$

$$for\ \ some\ \lambda \geq 0 \qquad\qquad w.r.t\ \beta$$

We now aim at solving the optimal $\hat{\beta}$ for any given value of $\lambda$. We need to derive two common matrix equations and show that;

i) $\frac{\partial(a^T\beta)}{\partial\beta} = a$ (18)

ii) $\frac{\partial(\beta^T A\beta)}{\partial\beta} = 2A\beta$ (19)

Where $a$ is a $2 \times 1$ vector, A is a $2 \times 2$ symmetric matrix, $\beta = [\beta_0\ \ \beta_1]^T$, and the partial $g(\beta)\quad w.r.t\quad \beta$ is;

$$\frac{\partial g(\beta)}{\partial\beta} = \begin{bmatrix}\partial g(\beta)/\partial\beta_0 \\ \partial g(\beta)/\partial\beta_1\end{bmatrix}$$

i) By multiplication we know that;

$$a^T\beta = a_1\beta_0 + a_2\beta_1$$

$$\therefore\ \frac{\partial(a^T\beta)}{\partial\beta} = \begin{bmatrix}\partial\ (a_1\beta_0 + a_2\beta_1)/\beta_0 \\ \partial\ (a_1\beta_0 + a_2\beta_1)/\beta_1\end{bmatrix} = \begin{bmatrix}a_1 \\ a_2\end{bmatrix} = a$$

ii) By multiplication we also know that;

$$A\beta = \begin{bmatrix}a_1 & a_2 \\ a_2 & a_3\end{bmatrix}\begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} = \begin{bmatrix}a_1\beta_0 & a_2\beta_1 \\ a_2\beta_0 & a_3\beta_1\end{bmatrix}$$

$$\beta^T A\beta = a_1\beta_0^2 + 2a_2\beta_0\beta_1 + a_3\beta_1^2$$

Using partial derivatives we obtain;

$$\frac{\partial(\beta^T A\beta)}{\partial\beta} = \begin{bmatrix}\dfrac{\partial(a_1\beta_0^2 + 2a_2\beta_0\beta_1 + a_3\beta_1^2)}{\beta_0} \\ \dfrac{\partial(a_1\beta_0^2 + 2a_2\beta_0\beta_1 + a_3\beta_1^2)}{\beta_1}\end{bmatrix}$$

$$= \begin{bmatrix}2a_1\beta_0 + 2a_2\beta_1 \\ 2a_2\beta_0 + 2a_3\beta_1\end{bmatrix}$$

$$= 2 \begin{bmatrix} a_1 & a_2 \\ a_2 & a_3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$= 2A\beta$$

### 3.2.2 Deriving the Penalized Spline Solution

The solution to the penalized spline will involve minimizing (17), that is solving when all the partial derivatives with respect to $\beta_0$ and $\beta_1$ are zero. This could be represented mathematically as;

$$\frac{\partial}{\partial \hat{\beta}} \left( \left\| y - X\hat{\beta} \right\|^2 \right) + \frac{\partial}{\partial \hat{\beta}} \left( \lambda^2 \hat{\beta}^T D \hat{\beta} \right) = 0 \qquad (20)$$

Since differentiation is linear, we are able to split (19) into two parts

With the two identities already proved we get;

$\frac{\partial}{\partial \hat{\beta}} \left( \left\| y - X\hat{\beta} \right\|^2 \right) = 2X^T(y - X\hat{\beta})$ where $X^T y$ is the vector $a^T$ and $X^T X$ is the matrix A. We also have;

$\frac{\partial}{\partial \hat{\beta}} \left( \lambda^2 \hat{\beta}^T D \hat{\beta} \right)$ which by linearity of differentiation, $\lambda$ gets factored out leaving;

$\lambda^2 \frac{\partial}{\partial \hat{\beta}} \left( \hat{\beta}^T D \hat{\beta} \right)$ with D being symmetrical, we gain apply the differentiation identities to get;

$\frac{\partial}{\partial \hat{\beta}} \left( \lambda^2 \hat{\beta}^T D \hat{\beta} \right) = 2\lambda^2 D\hat{\beta}$, we finally combine the partial derivatives to get;

$$\lambda^2 D\hat{\beta} - X^T \left( y - X\hat{\beta} \right) = 0 \qquad (21)$$

Clearly from linear algebra we can manipulate (20) to get;

$$\lambda^2 D\hat{\beta} = X^T \left( y - X\hat{\beta} \right)$$

$$\lambda^2 D\hat{\beta} = X^T y - X^T X\hat{\beta}$$

$$X^T X\hat{\beta} + \lambda^2 D\hat{\beta} = X^T y$$

$$\hat{\beta}(X^T X + \lambda^2 D) = X^T y$$

$$\hat{\beta} = (X^T X + \lambda^2 D)^{-1} X^T y$$

Now since we already have $\hat{\beta}$ and we know that $\hat{y} = X\hat{\beta}$ we now fit the penalized spline as follows;

$$\hat{y} = X(X^T X + \lambda^2 D)^{-1} X^T y \quad (22)$$

## 3.3 Smooth Hazard Model

### 3.3.1 Fitting the Penalized spline (P-Spline)

Given the survival time $\tau_i$ for the $ith$ observational unit, we define $C_i$ to represent the right censoring time; with $i = 1,2, \dots, N$. We note that $Y_i = \min(\tau_i, C_i)$. We also define the censoring indicator, $\delta_i$ as follows;

$$\delta_i = \begin{cases} 1 & if\ \tau_i < C_i \\ 0 & Otherwise \end{cases} \quad (23)$$

Now given a covariate $x_i$, which is independent of time and denoted by $p - dimensional$ covariate vector for the $ith$ observational unit, we can then model the hazard function as;

$$h(t, x_i) = \lambda_0(t)\exp\{(x_i^T \beta_x(t)\} \quad (24)$$

Where $\lambda_0$ is the baseline hazard, $\beta_x(t)$ is the vector of covariate effects that vary smoothly with survival time, $t$. The main idea is to estimate $\beta(t)$ smoothly by avoiding the tough parametric assumptions. A common approach to dealing with non-linear relationship is to approximate $f$ by a polynomial of order $m$ (Yuedong. W, 2011). For instance,

$$f(x) = \beta_0 + \beta_1 x + \cdots + \beta_{m-1} x^{m-1} \quad (25)$$

Applying the Sobolev Space, $f \in W_2^m[a, b]$ we have

$$W_2^m[a,b] = \left\{ f: f, f', \ldots, f^{m-1} \text{ are absolutely continuous}, \int_a^b \left(f^{(m)}\right)^2 dx < \infty \right\} \quad (26)$$

By Taylors theorem,

$$f(x) = \underbrace{\sum_{v=0}^{m-1} \frac{f^{(v)}(a)}{v!}(x-a)^v}_{Polynomial\ of\ order\ m} + \underbrace{\int_a^x \frac{(x-u)^{m-1}}{(m-1)!} f^{(m)}(u)du}_{Rem\ (x)} \quad (27)$$

The polynomial regression in (25) ignores the remainder term $Rem\ (x)$, it could be mere assumption that $Rem\ (x)$ is negligible. The idea behind smoothing spline is simply to let data decide how large $Rem\ (x)$ is going to be. Now using the least squares (LS) on $W_2^m[a,b]$, an infinite dimensional space, we have;

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i)^2 \quad (28)$$

The distance measure between $f$ and polynomial is,

$$\int_a^b \left(f^{(m)}\right)^2 dx \quad (29)$$

We now estimate $f$ by minimizing $LS$ under the constraint say, $\rho$ which yields;

$$\int_a^b \left(f^{(m)}\right)^2 dx \leq \rho \quad where\ \rho\ is\ some\ constant.$$

We introduce a Lagrange multiplier in (28) and (29) so as to get Penalized Least Squares (PLS)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i)^2 + \lambda \int_a^b \left(f^{(m)}\right)^2 dx \quad (30)$$

$\int_a^b \left(f^{(m)}\right)^2 dx$ is called the roughness penalty

If we consider now the Sobolev space $W_2^m[a,b]$ with a linear product

$$(f,g) = \sum_{v=0}^{m-1} f^{(v)}(a)g^{(v)}(a) + \int_a^b f^{(m)}\, g^{(m)} dx \tag{31}$$

We further say that $W_2^m[a,b] = \mathcal{H}_0 \oplus \mathcal{H}_1, where$

$$\mathcal{H}_0 = span\,\{1,(x-a),\dots,(x-a)^{m-1}/(m-1)!\} \tag{32}$$

$$\mathcal{H}_1 = \left\{ f: f^{(v)}(a) = 0, v = 0,\dots,m-1, \int_a^b \left(f^{(m)}\right)^2 dx < \infty \right\} \tag{33}$$

Now (31) and (32) are RKHS's with the RKs

$$R_0(x,z) = \sum_{v=1}^{m} \frac{(x-a)^{v-1}(z-a)^{v-1}}{(v-1)!\,(v-1)!} \tag{34}$$

$$R_1(x,z) = \int_a^b \frac{(x-u)_+^{m-1}(z-u)_+^{m-1}}{(m-1)!\,(m-1)!}\, du \tag{35}$$

$(x)_+$ means that $max(x,0)$

Looking at (31), it is clear that $\mathcal{H}_0$ contains a polynomial of order $m$ in the Taylor expansion. If we now denote $Q$ to be the orthonormal projection operator onto $\mathcal{H}_1$ and based on the definition of the inner product, the roughness penalty is;

$\int_a^b \left(f^{(m)}\right)^2 dx = \|Qf\|^2$ which shows that $\int_a^b \left(f^{(m)}\right)^2 dx$ measures the distance between parametric polynomial space $\mathcal{H}_0\ and\ f.\mathcal{H}_0$ has no penalized functions. The penalized least squares is thus;

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x))^2 + \lambda\|Qf\|^2$$

Where $\lambda$ is a smoothing parameter that controls the balance between the goodness-of-fit measured by the least squares and departure from the null space $\mathcal{H}_0$ measured by $\|Qf\|^2$. Functions in $\mathcal{H}_0$ are not penalized since

$\|Qf\|^2 = 0$ when $f \in \mathcal{H}_0$.

## 3.4 Determination of the smoothing parameter $\lambda$

Residual sum of squares has been proposed by Griggs (2013) to be a good measure of the "goodness-of-fit" since its summation obtains the overall error between the actual data and the regression curve. The residual sum of squares (RSS) is defined as;

$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

RSS, however, faces challenges in fitting the penalized spline model with many knots and as such we propose Cross-Validation method.

### 3.4.1 Cross-Validation method

Cross-Validation (CV) is used to assess the fit of the model with $\lambda$ in a similar way as RSS. The CV however, removes the $y_i$ point and evaluates how well the given fit predicts the removed point. CV, attempts to minimize RSS while assuming the closest point. Since the method removes the $y_i$ point it is sometimes referred to as "leave-one-out" approach. The strategy is defined as;

$$CV(\lambda) = \sum_{i=1}^{n}\{y_i - \hat{f}_{-1}(x_i;\lambda)\}^2$$

Where $\hat{f}_{-1}(x_i;\lambda)$ is the spline fit lacking $(x_i, y_i)$ point. This allows us to obtain the value λ for any given spline basis minimizing this value while taking into account the

prediction of the new points and avoiding over-fitting. The method is however computationally intensive hence we use an approach given below;

$$\hat{f}_{-1}(x_i;\lambda) = \frac{\sum_{i=1,i\neq j}^{n} S_{\lambda,ij} y_j}{\sum_{i=1,i\neq j}^{n} S_{\lambda,ij}}$$

Where $S_\lambda$ refers to the smoothing matrix of the penalized linear spline (i.e. $S_\lambda = \mathrm{X}(X^T\mathrm{X} + \lambda^2\mathrm{D})^{-1}X^T$). The CV thus can be rewritten as;

$$CV(\lambda) = \sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - S_{\lambda,ii}}\right)^2$$

The above equation significantly reduces the computational time since it applies the normal residual of the previously fitted model that only requires the diagonal entries of the smoothing matrix.

## 3.5 **Asymptotic properties of the penalized splines**

In investigating the properties of the penalized spline estimator, we look at the average mean squared error (AMSE) and the asymptotic bias and variance of the model. We also discuss the optimum choice of the smoothing parameter $\lambda$.

### 3.5.1 Average Mean Squared Error (AMSE)

According to Demmler and Reinsch (1975), it is possible to express the average bias and variance in terms of the eigenvalues having been obtained from the singular value decomposition.

$$(N^tN)^{-t/2}D_q(N^tN)^{-1/2} = Udiag(s)U^t$$

Such that $U$ is the eigenvectors matrix and $s$ represents the eigenvalues$s_j$. We denote $A = N(N^tN)^{-1/2}U$, with matrix $A$ being the semi-orthogonal with $A^tA = I_{k+p+1}$ and $AA^t = N(N^tN)^{-1}N^t$. The penalized spline estimator can be rewritten as;

$$\hat{f} = A\{I_n + \lambda diag(s)\}^{-1}A^tY = \{I_n + \lambda diag(s)\}^{-1}AA^tY$$
$$= \{I_n + \lambda diag(s)\}^{-1}\hat{f}_{reg} - f$$

We can now obtain the AMSE

$$AMSE(\hat{f}) = \frac{1}{n}E\left\{(\hat{f} - f)^t(\hat{f} - f)\right\}$$
$$= \frac{\sigma^2}{n}\sum_{j=1}^{k+p+1}\frac{1}{(1+\lambda s_j)^2} + \frac{\lambda^2}{n}\sum_{j=1}^{k+p+1}\frac{s_j^2 b_j^2}{(1+\lambda s_j)^2} + \frac{1}{n}f^t(I_n - AA^t)f$$

$AA^t$ is an idempotent matrix and $AA^tf = E(\hat{f}_{reg})$ also $f = \{f(x_1), \dots, f(x_n)\}^t$ and $b = A^tf$ thus

$$AMSE(\hat{f}) = \frac{\sigma^2}{n}\sum_{j=1}^{k+p+1}\frac{1}{(1+\lambda s_j)^2} + \frac{\lambda^2}{n}\sum_{j=1}^{k+p+1}\frac{s_j^2 b_j^2}{(1+\lambda s_j)^2}$$
$$+ \frac{1}{n}\sum_{j=1}^{n}\left[E\{\hat{f}_{reg}(x_j)\} - f(x_j)\right]^2$$

In the above equation the first term is the average variance while the second term is the average squared bias (shrinkage) and the third term is the average squared approximation bias. We define $K_q = (K + p + 1 - q)\,(\lambda n)^{1/(2q)_n - 1/(2q)}$ and consider two asymptotic scenarios.

i) When $K_q < 1$ and $f(.) \in C^{p+1}[a, b]$ we have

$$AMSE\,(\hat{f}) = O\left(\frac{K}{n}\right) + O\left(\frac{\lambda^2}{n^2}K^{2q}\right) + O\left(K^{-2(p+1)}\right)$$

ii) When $K_q \geq 1$ and $f(.) \in W^q[a, b]$ we have

$$AMSE\,(\hat{f}) = O\left(\frac{n^{1/(2q)-1}}{\lambda^{1/(2q)}}\right) + O\left(\frac{\lambda}{n}\right) + O(K^{-2q})$$

We observe that when $K_q < 1$ the result obtained is similar to the regression splines. AMSE is determined by the squared approximation bias and the mean asymptotic variance. The smaller the smoothing parameter, $\lambda$, the negligible the shrinkage bias is. We also observe that when $K_q \geq 1$ the result obtained is similar to the smoothing spline. AMSE in this case is dominated by the squared shrinkage bias and mean

asymptotic variance. The mean squared approximation bias has the same order as that of the shrinkage bias when $K_q = 1$ and is negligible when $K_q > 1$. The results suggest that convergence rate for the penalized spline estimator is much faster when $K_q < 1$ basing on the assumption that $q \leq p$.

### 3.5.2 Asymptotic variance and Bias

We derive the asymptotic variance and bias considering the above two asymptotic scenarios.

i) When $K_q < 1$ and $f(.) \in C^{p+1}[a, b]$ we have

$$Var\left\{\hat{f}(x)\right\} = \frac{\sigma^2}{n} N(x)\left(G + \lambda D_q/n\right)^{-1} G\left(G + \lambda D_q/n\right)^{-1} N^t(x) + 0\left\{(n\delta)^{-1}\right\}$$

$$E\left\{\hat{f}(x)\right\} - f(x) = b_a(x; p+1) + b_\lambda(x) - o(\delta^{p+1}) + o\left(\lambda n^{-1}\delta^{-q}\right)$$

ii) When $K_q \geq 1$ and $f(.) \in W^q[a, b]$ we have

$$Var\left\{\hat{f}(x)\right\} = \frac{\sigma^2}{n} N(x)\left(G + \lambda D_q/n\right)^{-1} G\left(G + \lambda D_q/n\right)^{-1} N^t(x) + 0\left\{\left(n^{-1}\left(\lambda/n\right)^{-1}\right)^{-1/2q}\right\}$$

$$E\left\{\hat{f}(x)\right\} - f(x) = b_a(x; q) + b_\lambda(x) + o(\delta^q) + o\left\{\left(\lambda/n\right)^{1/2}\right\}$$

# CHAPTER 4

# RESULTS

## 4.0 Introduction

In this chapter we explore and identify non-linear and time varying effects in survival data. We also describe the data set used in this study.

## 4.1 Data

Data was obtained from Nairobi Hospital-Cancer Registry. The patients with primary node-positive breast cancer registered between 2008 and 2012. The breast cancer data contains binary and continuous variables. Complete data for the prognostic factors age, tumour size, tumour grade, Menopause, number of positive lymph nodes, progesterone and estrogen receptor concentration available for 277 patients was analyzed.

**Table 1: Covariates in the breast cancer data**

| Covariate | Median or Percent |
|---|---|
| Age (years) | 52 |
| Menopausal status | 46% Pre, 54% Post |
| Progesterone receptor | 135 |
| Estrogen receptor | 77 |
| Hormonal therapy | 64% No, 36% Yes |
| Tumor size | 25 |
| Tumor grade | 21% 1, 70% 2, 9% 3 |

## 4.2 Assessment of proportional hazards model

The Cox model specifies that hazard function for the failure time $T$ commonly associated with a $1 \times q$ column covariate vector $Z$ takes the form;

$\lambda(t; Z) = \lambda_0(t) \exp(\beta' Z)$

Where $\lambda_0(t)$ represents the baseline hazard function while $\beta'$ is a $1 \times q$ column vector of the covariates.

(Lin, D. Y. & Wei, L. J, 1991), found out that both numerical and graphical methods can be used to assess the model based on cumulative sum of martingale transformations and residuals. The distribution of such stochastic processes under the assumed model could be approximated using the distributions of certain zero-mean Gaussian processes whose realizations can only be generated by simulation. We compare the observed residual pattern both numerically and graphically. The comparisons enable us to assess objectively whether or not the observed residual pattern has a reflection of anything beyond random fluctuation. The procedures are essential when it comes to determination of the appropriate functional forms of covariates and in the assessment of the proportional hazards assumption.

Now, let us consider a sample of $n$ subjects and letting $(X_i, \delta_i, Z_i)$ to be the data of the $ith$ subject for instance, $X_i$ representing the observed failure time, $\delta_i$ representing the censoring indicator $\delta_i = 1$ if $T_i < C_i$ and $\delta_i = 0$otherwise, and $Z_i = \left(Z_{1i}, \dots, Z_{pi}\right)'$ is a $1 \times q$ vector of covariates. Let $N_i(t) = \delta_i I(X_i \leq t)$ and $Y_i(t) = I(X_i \geq t)$. We further let,

$S^0(\beta, t) = \sum_{i=1}^{n} Y_i(t) \exp(\beta' Z_i)$ and $Z(\beta, t) = \frac{\sum_{i=1}^{n} Y_i(t) \exp(\beta' Z_i) Z_i}{S^0(\beta, t)}$

Let $\hat{\beta}$ be the maximum partial likelihood estimate of , and we also let $\xi(\hat{\beta})$ to be the observed information matrix. We then define martingale residual as;

$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(u) \exp(\beta' Z_i) d\widehat{\Lambda}_0(u), 1, 2, 3, \dots, n$ where $\widehat{\Lambda}_0(u) = \int_0^t \frac{\sum_{i=1}^{n} dN_i(u)}{S^0(\hat{\beta}, t)}$

The empirical score process $U\left(\hat{\beta}, t\right) = U_1\left(\hat{\beta}, t\right)$ is a transformation of the martingale residuals given as;

$$U(\hat{\beta}, t) = \sum_{i=1}^{n} Z_i \, \widehat{M}_i(t)$$

### 4.2.1 Testing for the PH assumption

#### 4.2.1.1 Graphical Method

If the dataset comprises of categorical variables then it is possible to use Kaplan-Meier plot for survival distribution, that is , we plot for each level of covariate. In the presence of PH the curves steadily drift apart (parallel curves). Transformation of the Kaplan-Meier survival curves can also be applied and in such a case we plot the $log(-logS(t))$ as a function of the log survival time and just like the KM plots, the stratum log-minus plots exhibits constant differences (curves that are close to parallel).

These particular visualization methods are simple and easy to implement, they however a number of limitations. For instance, if the covariate has more than two categorical levels then KM plot is rendered useless in discerning presence or absence of PH since the graphs become much cluttered (Therneau, T. M. & P. M. Grambsch, 2000). Many a times, the log-minus plots are also hardly parallel, hence they tend to be less precise. It is therefore subjective to accept the PH hypothesis since one has to have very strong evidence to either conclude that the PH assumption has been violated or not. We now test for the PH assumption for the prognostic factors in the breast cancer data.

Figure 1 displays the plot for menopause and we see in the first figure (stphplot for meno) shows nonparallel lines, implying that the proportional-hazards assumption for the menopause has been violated. This is confirmed in the second figure (stcoxkm for meno), where the observed values and predicted values are somehow far apart.

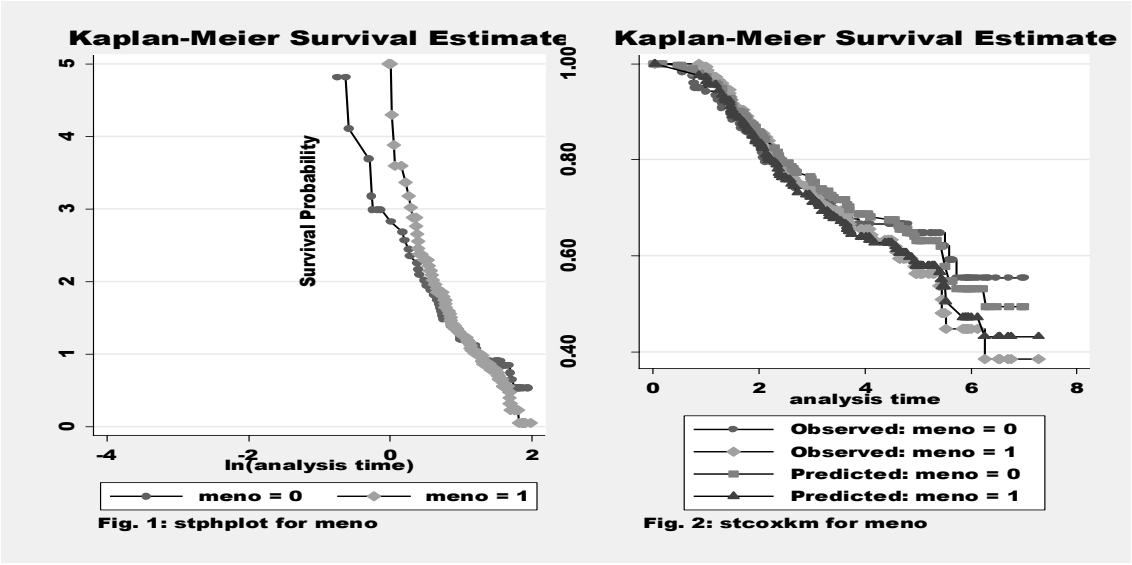


**Figure 1: Checking Proportional Hazard assumption for menopause**

For the grade, we observe a display of parallel lines in the first figure (stphplot for grade), suggesting that the proportional-hazards assumption for the grade has not been violated. This is confirmed in the second figure (stcoxkm for meno), where the observed values and predicted values are close together.

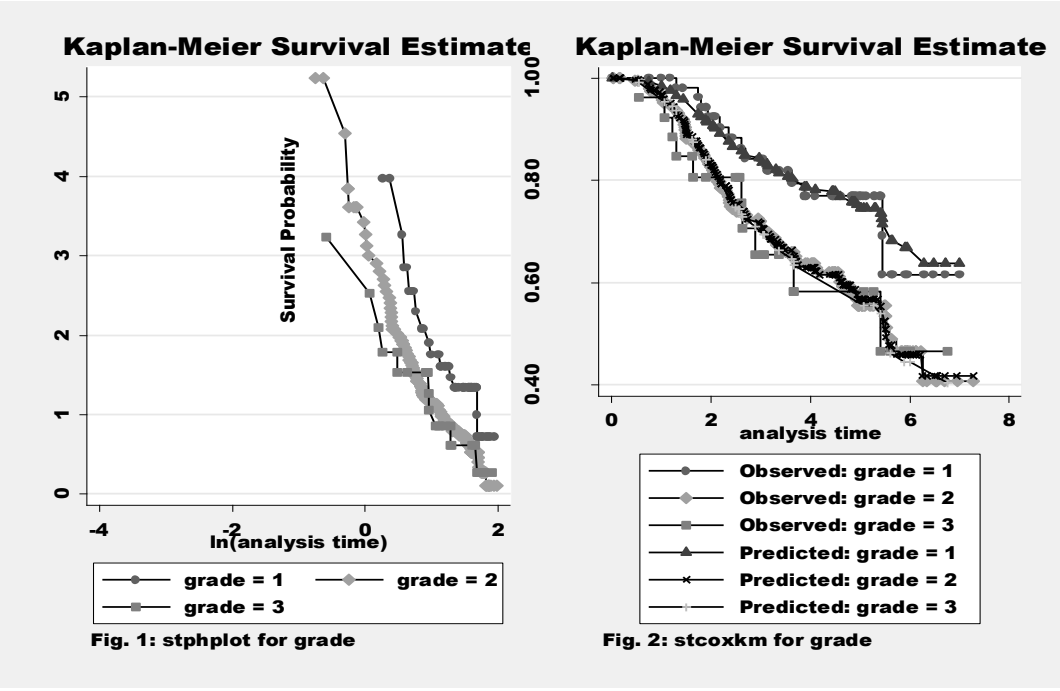


**Figure 2: Checking Proportional Hazard assumption for grade**

In figure 3 below, we test the PH assumption of the hormone variable. The stphplot displays lines that are parallel to each other, a clear indication that the proportional-hazards (PH) assumption for hormone has not been violated. This is confirmed by the stcoxkm plot where the observed values and predicted values are close together.

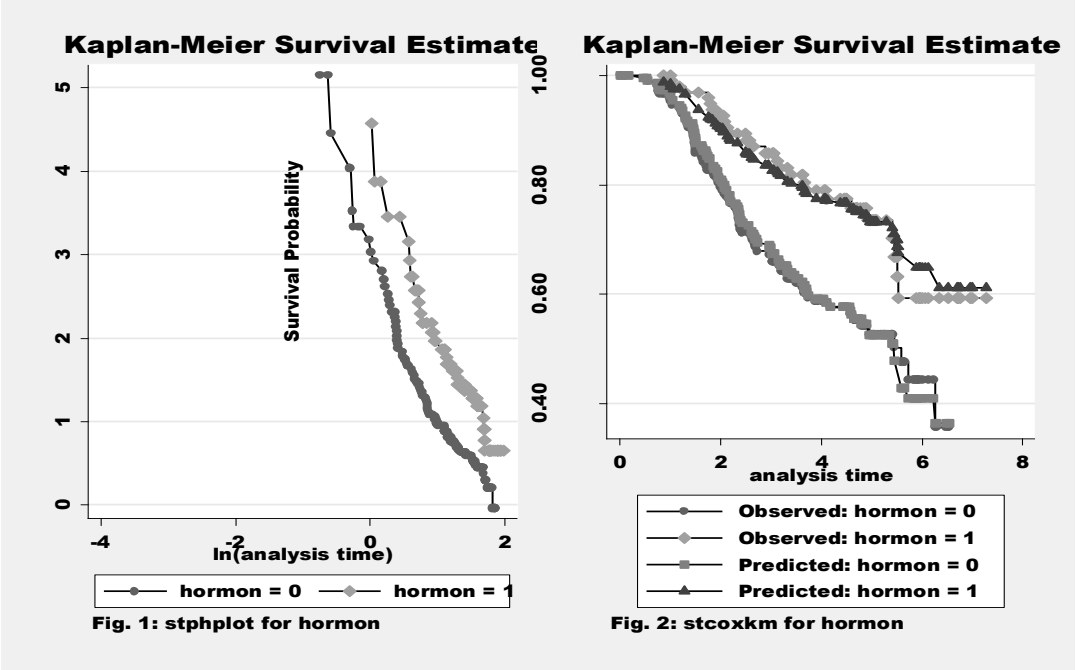


**Figure 3: Checking Proportional Hazard assumption for hormone**

Figure 4, tests the PH assumption of the gradd1 variable. Just like in the hormone case, the stphplot displays lines that are parallel to each other, a clear indication that the proportional-hazards (PH) assumption for gradd1 has not been violated. This is confirmed by the stcoxkm plot where the observed values and predicted values are close together.

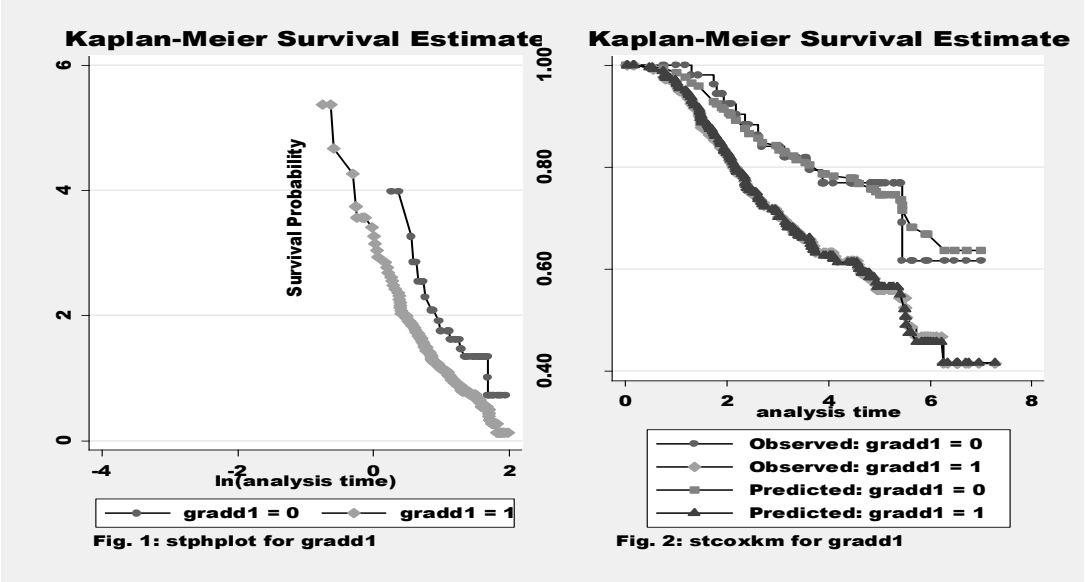


**Figure 4: Checking Proportional Hazard assumption for gradd1**

In the case of gradd2, we clearly observe that the PH assumption has been violated since the stphplot displays lines crossing each other (nonparallel lines). This has further been confirmed by the stcoxkm plot where the observed values and predicted values are close together

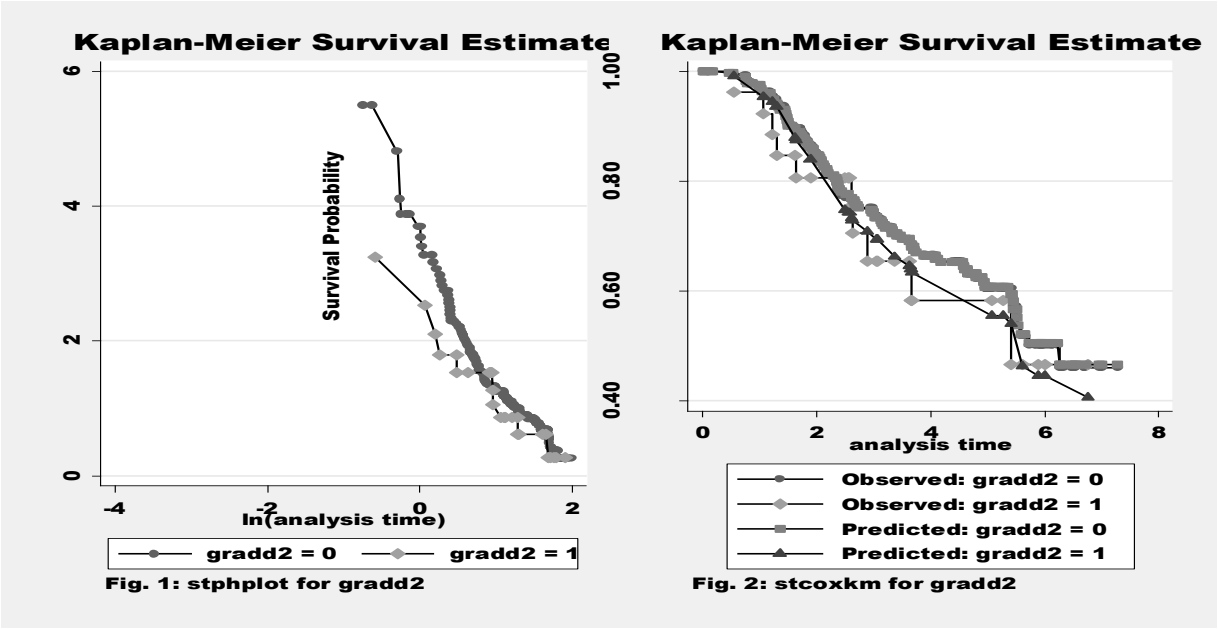


**Figure 5: Checking Proportional Hazard assumption for gradd2**

#### 4.2.1.2 Score test based on scaled schoenfeld residuals

We define schoenfeld residuals after running a Cox model for each predictor variable in the model. That is to say that the number of Schoenfeld residual variables in the model is the same as the number of covariates. The residuals are associated with the contributions of each of the covariates to the log partial likelihood. Scaled Schoenfeld residuals can be very useful in diagnostics of Cox regression models, more so in the assessment of the PH assumption (P. M. Grambsch & T. M. Therneau, 1994). In theory, the scaled Schoenfeld residuals are the adjusted Schoenfeld residuals based on the inverse of the covariance matrix of the Schoenfeld residuals. Grambsch and Therneau (1994) suggest that under the assumption that that the distribution of the predictor variable is similar in the various risk sets, the adjustment can be performed using the variance-covariance matrix of the parameter estimates divided by the number of events in the sample. The null hypothesis for the test on proportional hazards based on the scaled Schoenfeld residuals is that the slope of Schoenfeld residuals against a function of time is zero for each covariate variable. Once the scaled Schoenfeld residuals are created, we can then perform the test using generalized linear regression approach. Specifically, the test statistic on an individual covariate is;

$$\frac{\left[\sum_{i=1}^{N}\{\delta_i g(t_i)-\bar{g}(t)\}r_s\right]^2}{\Delta\hat{V}_{uu}\sum_{i=1}^{N}(\delta_i g(t_i)-\bar{g}(t))^2}$$

In this formula, $r_s$ represents the variable of scaled Schoenfeld residuals, $g(t)$ on the other hand is the predefined function of time set before the test, $\delta_i$ represents the indicator variable of event, $\Delta$ represents the total number of events and $V_{uu}$ represents

the estimate for the variance of the parameter estimate of the covariate of interest. The sum is taken over all the observations in the data. The test statistic is asymptotically distributed as a $\chi^2$ having 1 degree of freedom. The overall test statistic for the $p$ predictor variables is given as follows.

$$\left[\sum_{i=1}^{N}\{\delta_i g(t_i) - \bar{g}(t)\} r_s\right]' \left[\frac{\Delta\hat{V}}{[\sum_{i=1}^{N}\{\delta_i g(t_i) - \bar{g}(t)\}]^2}\right]\left[\sum_{i=1}^{N}\{\delta_i g(t_i) - \bar{g}(t)\} r_i\right] \quad (36)$$

In the formula, $r_i$ represents the vector of the unscaled Schoenfeld residuals of interest. It has $p$ degrees of freedom and it asymptotically follows a $\chi^2$ distribution.

We fit a new Cox model and perform the test for proportional hazards:

**Table 2: Cox model**

```
. stcox age meno size grade gradd1 gradd2 nodes enodes pgr er hormon, nolog

         failure _d:  1 (meaning all fail)
   analysis time _t:  id
note: gradd2 omitted because of collinearity

Cox regression -- Breslow method for ties

No. of subjects =          277                     Number of obs   =       277
No. of failures =           98
Time at risk    =  977.5222414
                                                   LR chi2(10)     =     46.77
Log likelihood  =    -476.2825                     Prob > chi2     =    0.0000
```

| _t | Haz. Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | .9746729 | .016991 | -1.47 | 0.141 | .9419337 | 1.00855 |
| meno | 1.856089 | .5836122 | 1.97 | 0.049 | 1.0022 | 3.437503 |
| size | 1.016248 | .0058679 | 2.79 | 0.005 | 1.004812 | 1.027814 |
| grade | .9691075 | .3363496 | -0.09 | 0.928 | .4908448 | 1.913373 |
| gradd1 | 1.916974 | .9406846 | 1.33 | 0.185 | .73269 | 5.015476 |
| gradd2 | 1 | (omitted) | | | | |
| nodes | .9982625 | .0343654 | -0.05 | 0.960 | .9331296 | 1.067942 |
| enodes | .2230905 | .1979079 | -1.69 | 0.091 | .039207 | 1.2694 |
| pgr | .99812 | .0009922 | -1.89 | 0.058 | .9961773 | 1.000067 |
| er | 1.000628 | .0008388 | 0.75 | 0.454 | .9989856 | 1.002274 |
| hormon | .4190189 | .1025325 | -3.55 | 0.000 | .2593867 | .6768924 |

**Table 3: Test for proportional hazard (PH) assumption**

```
. estat phtest, detail
Test of proportional-hazards assumption
    Time:  Time
------------------------------------------------------------------------------------------
                |     rho           chi2       df       Prob>chi2
------------+-----------------------------------------------------------------------------
age             |    0.14069        2.48        1         0.1154
meno            |    0.04155       10.19        1         0.0213
size            |   -0.16176        1.73        1         0.1885
grade           |    0.01940        0.04        1         0.8480
gradd1          |   -0.06132        0.39        1         0.5311
gradd2          |    0.04341       11.23        1         0.0321
nodes           |    0.10745        0.92        1         0.3380
enodes          |    0.06570        0.37        1         0.5429
pgr             |   -0.02293        0.06        1         0.7997
hormon          |    0.11692        1.29        1         0.2552
------------+-----------------------------------------------------------------------------
global test |                     19.10        9         0.0400
------------------------------------------------------------------------------------------
```

From table 3 above, we clearly observe that two variables (menopause and gradd2) violate the proportional hazards assumption.

The test for the individual predictors uses the unscaled Schoenfeld residuals, while the global test uses the scaled Schoenfeld residuals (Therneau, T. M. & P. M. Grambsch, 2000).

## 4.3 Testing for the time variation

In 1994, Hess proposed a test based and graphical approaches for exploring possible violations of Proportional Hazard assumption. A simple but informal method is by

estimating if the constant estimator of the PH-model lies in between the standard error (SE) bands of the dynamic estimation.

A part from the graphical methods tests for the goodness-of-fit for the null hypothesis, $H_0: \beta(t) = \beta$ can also be used. Based on the Cox proposal, the goodness-of-fit is estimated by modifying the predictor $\eta(t) = (\beta_0 + \varphi(t)\beta_1)X$ of the following model;

$\lambda(t|X) = \lambda_0(t) \exp\{(\beta_0 + \varphi(t)\beta_1)X\}$ and we test the null hypothesis $H_0: \beta_1 = 0$ using the likelihood-ratio statistic.

Schonfeld (1980) suggested omnibus goodness-of-fit tests that compare the observed and expected frequencies of failure for a particular partition of time. It has been noted that the cubic regression spline approach that are based on fixed knots also allows for the formal testing of the proportional Hazard assumption.

## 4.4 Assessment of time-varying effects

### 4.4.1 Aalen Linear Hazards Model (ALHM)

In his paper, (Aalen, 1980) proposed a generalized linear model for the estimation of time varying effects based on regression coefficients. Issues of assessment, testing and estimation of the model fit have been discussed in Aalen (1989 & 1993). The model is given as;

$$h\big(t, x, \beta(t)\big) = \beta^0(t) + \beta^1(t)x^1 + \beta^2(t)x^2 + k + \beta_p(t)x_p \qquad (37)$$

The coefficients in the model vary with time, $t$. In checking for possible time-varying effects of the covariates, we use cumulated regression coefficients. The cumulative hazard function is obtained through integration of the hazard function and this yields;

$$H(t, x, \beta(t)) = \int_0^t h\,(u, x, \beta(u))du = \sum_{k=0}^{p} x_k \int_0^t \beta_k(u)du = \sum_{k=0}^{p} x_k\, B_k(t) \qquad (38)$$

$B_k(t)$ is the cumulative regression coefficient for the $kth$ covariate and $B_0(t)$ is the baseline cumulative hazard function.

### 4.4.2 Testing for time-varying effects

We tested for the time varying effects using the Aalen model. Our hypothesis is;

$H_0: B_k(t) = 0$ for $k = 0,1,2,\ldots,p$. The results from the test portion are shown below;

```
. stlh age meno size grade gradd1 gradd2 nodes enodes pgr hormon, test(1 2 3 4)
>  nograph
note: gradd1 dropped because of collinearity

Graphs and tests for Aalen's Additive Model
-------------------------------------------
Model:  age meno size grade gradd2 nodes enodes pgr hormon
Obs:    277

..

Test 1: Uses Weights Equal to
1.0

Variable         z          P
------------------------------
age            0.060      0.952
meno          -0.049      0.961
size           0.750      0.453
grade          0.664      0.507
gradd2        -1.029      0.303
nodes          1.070      0.285
enodes         0.973      0.331
pgr           -0.039      0.969
hormon        -2.291      0.022
_cons         -0.898      0.369

Test 2: Uses Weights Equal to
the Size of the Risk Set

Variable         z          P
------------------------------
age           -1.138      0.255
meno           1.736      0.083
size           2.468      0.014
grade          2.225      0.026
gradd2        -1.165      0.244
nodes          1.154      0.249
enodes         0.637      0.524
pgr           -1.736      0.083
hormon        -3.570      0.000
_cons         -0.379      0.705

Test 3: Uses Weights Equal to
Kaplan-Meier Estimator at Time t-

Variable         z          P
------------------------------
age           -0.148      0.882
meno           0.230      0.818
size           1.277      0.202
grade          1.057      0.291
gradd2        -1.084      0.279
nodes          1.097      0.273
enodes         0.963      0.336
pgr           -0.342      0.733
hormon        -2.758      0.006
_cons         -0.876      0.381

Test 4: Uses Weights Equal to
(Kaplan-Meier Estimator at Time t-)/(Std. Dev of the Time-varying Coefficient)

Variable         z          P
------------------------------
age           -0.044      0.965
meno           2.993      0.003
size           1.120      0.263
grade          6.184      0.000
gradd2        -5.093      0.000
nodes         -2.194      0.028
enodes        -1.654      0.098
pgr           -3.448      0.001
hormon        -4.885      0.000
_cons          1.338      0.181
```

Testing using weights equal to one yields only one significant test for the covariate effects. However, when we test using weights equal to the size of the risk set, we obtain three significant tests for the covariate effects. Test results using weights equal to the product of the Kaplan-Meier or just Kaplan-Meier weights and the inverse of

the estimated standard deviation of B(t) shows that only one test is significant for the covariate effects.

## 4.5 Simulation

Simulation studies have been used in evaluating the performance and properties of the statistical models (Burton, Altman, Royston, & Holder, 2006). In the field of survival analysis either weibull distribution (that makes an assumption of a monotonically decreasing or increasing hazard) or the exponential distribution (that makes an assumption of constant hazard function) are implemented and used. In the analysis of cancer data, a turning point is in most cases observed in the hazard function.

We simulated survival times from a penalized model with an increasing hazard ratio. We first generated a binary treatment group indicator by incorporating a constant treatment effect. The seed for reproducibility was set and we conducted 100, 250, 500, 1000 and 5000 replicates in order to analyze the bias and convergence of the estimate of the treatment effect. The model was defined as;

$$h(t) = \lambda\gamma t^{\gamma-1} exp(\beta X_i + \phi X_i \times \log(t))$$

We simulated a single data set and fitted a flexible (penalized) parametric survival model that allowed for time-dependent hazard ratio for the effect of treatment.

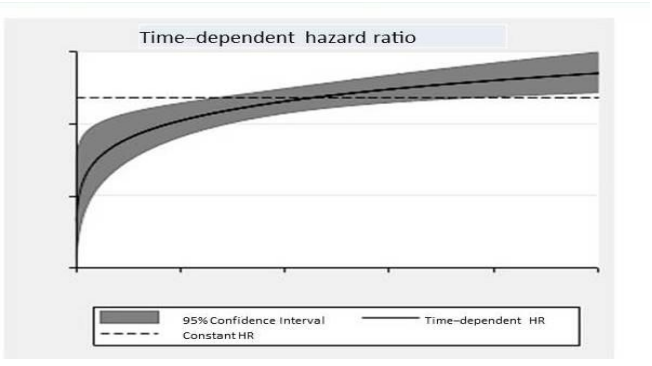


**Figure 6: Time-dependent hazard ratio**

### 4.5.1 Numerical Results

In this section we applied simulation data in order to check on the asymptotic properties as well as the performance of the penalized spline model. Using $x_i$ as the independent variable generated from a uniform distribution on the interval [0, 1] we created the dependent variable $Y_i$ such that $Y_i = Sin(2\pi x) + \varepsilon_i$ where $\varepsilon_i$

Is the error term distributed as;

i) Normally distributed with mean zero and variance 0.05.

ii) Exponentially distributed with mean 2.

iii) Having a Cauchy distribution with scale 0.02 and location 0.

**Table 4: Results of AMSE for n=100, 250, 500 and 1000**

| n=100 | Normal | Exponential | Cauchy |
|---|---|---|---|
| 0.01 | 11.56 | 5.1 | 4647.26 |
| 0.1 | 4.56 | 8.34 | 378.21 |
| 0.25 | 3.96 | 12.56 | 25.67 |
| 0.5 | 3.01 | 24.3 | 24.91 |
| n=250 | Normal | Exponential | Cauchy |
| 0.01 | 5.36 | 3.41 | 2567.32 |
| 0.1 | 2.43 | 4.01 | 56.78 |
| 0.25 | 1.96 | 7.56 | 3.67 |
| 0.5 | 1..47 | 14.5 | 2.97 |
| n=500 | Normal | Exponential | Cauchy |
| 0.01 | 3.65 | 1.47 | 652.1 |
| 0.1 | 1.78 | 2.01 | 29.65 |
| 0.25 | 1.41 | 4.25 | 1.56 |
| 0.5 | 1.23 | 5.32 | 0.93 |
| n=1000 | Normal | Exponential | Cauchy |
| 0.01 | 2.43 | 0.75 | 257.49 |
| 0.1 | 0.98 | 1.86 | 17.45 |
| 0.25 | 0.45 | 2.89 | 0.97 |
| 0.5 | 0.17 | 3.21 | 0.22 |

Table 4 shows the results for AMSE based on different values of $\tau$ that is, when $\tau$=0.5, 0.25, 0.1 and 0.01. The quantile performance of the penalized spline model with a normally distributed error term is good at all values of $\tau$. However, the AMSE with the Cauchy distribution is very small $\tau$=0.5 implying presence of a robust estimator.

# CHAPTER 5

## DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

### 5.1 Discussion

In this paper we propose the use of penalized splines in order to check, detect and model the time-varying effects in survival data within the context of Cox PH framework. The model allows for formal testing of the time-varying effects using standard methods. In our simulation study, we observe consistency with a high degree of checking and detecting time-variation. The graphical test for the PH assumption is subject to bias and is very difficult to use when a categorical prognostic factor with many values is given.

Penalized spline model provides a useful tool for the analysis of survival data with no pre-defined information on the time-varying effect and where the PH assumption seems to be doubtful.

### 5.2 Conclusions

In the analysis of larger studies of censored data with long term follow-up, the usual common standard techniques such as Cox model (Cox, 1972) may not be appropriate due to violation of the proportional hazard assumption that is caused by the time-varying effects. By ignoring the presence of such time-varying effects one may end up with incorrect models coupled with biased conclusions as a result of misleading effect estimates. Appropriate modeling of the shapes of the covariates is very important since ‘incorrect’ shapes of the time varying effects could result to misleading conclusions just as erroneously assuming the proportional hazard. Previous studies have shown varying tests and models for the time-varying effects. Cox (1972) proposed a transformation of time which formed a basis for testing and

assessing the non-PH, a method that heavily relies on the choice of the time transformation. In this paper we proposed the use of penalized splines in order to disclose and model effects of survival data within the context of cox model framework. The model allows for easy testing of time variation in the presence of effects using standard methods such as likelihood ratio test. However, although the penalized splines (PS) provide a flexible fit, they still suffer from the same restrictions that affect other non-linear smooth functions such as Fractional Polynomials.

## 5.3 Recommendations

Several advanced techniques such as fractional polynomial (FP) and restricted cubic spline (RCS) have been proposed. Many of such models have however experienced both technical and theoretical setbacks. The approaches have for instance fallen short of building multivariable model strategies for the selection of time-varying effects. Future research should thus employ a larger simulation data to study the properties of the various approaches and provide comparisons of the different techniques for selection and modeling time-varying effects.

# REFERENCES


Aalen. O. O. (1980). A model for non-parametric regression analysis of counting processes. *Lecture Notes on Statistics 2*.

Aalen. O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine 8 (8),* 907–25.

Aalen. O. O. (1993). Further results on the non-parametric linear regression model in survival analysis. *Statistics in Medicine 12 (17),,* 1569–88.

Abrahamowicz, M., , T. MacKenzie,, & J. M. Esdaile. (1996). Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association 91,* , 1432–1439.

Berger, U., , J. Schäfer, , & K. Ulm. (2003). Dynamic Cox modelling based on fractional polynomials: Time–variations in gastric cancer prognosis. . *Statistics in Medicine 22(7),* , 1163–1180.

Boor, d. C. (1978). A Practical Guide to Splines. *Springer-Verlag: New-York.*

Brown, D., , G. Kauermann, , & I. Ford. (2007). A partial likelihood approach to smooth estimation of dynamic covariate effects using penalised splines. . *Biometrical Journal 49,,* 1–12.

Buchholz, A. (2010). Assessment of time–varying long–term effects of therapies and prognostic factors.

Cai, Z. , & Y. Sun. (2003). Local linear estimation for time-dependent coefficients in cox's regression model. . *Scandinavian Journal of Statistics 30(1),,* 93–111.

Coradini, D., et.al. (2000). Time-dependent relevance of steroid receptors in breast cancer. *Journal of Clinical Oncology 18,,* 2702–2709.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological 34,* , 187–220.

Eilers, P. H. C. , & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Stat. Science 11 (2),* , 89-121.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics 19: 1.* .

Gao C, , Yang M,, Wu Y, , & et al. (2006). Hybrid coronary revascularization by endoscopic robotic coronary artery bypass grafting on beating heart and stent placement. *Ann Thorac Surg, 87,* 737-41.

Gore, S.M, Pocock, S.J., & Kerr, G.R. (1984). Regression models and non-proportional hazards in the analysis of breast cancer survival. . *Applied Statistics, 33(2),* , 176–195.

Gray. (1994). Spline based tests in survival analysis. *Biometrics*, 640–652.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. . *Journal of the American Statistical Association 87,* , 942–951.

Green, D. J. , & Silverman, B. W. (1994). onparametric Regression and generalized linear models. *N Chapman & Hall.*

Hastie, T, & Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazard model. *Biometrics 46,* 1005-1016.

Hastie, T. , & Tibshirani, R. (1993). Varying coefficient models. *Journal of the Royal Statistical Society, Series B 55,* , 757-796.

Heinzl, H. , & A. Kaider. (1997). Gaining more flexibility in cox proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine 54(3),* , 201–208.

Hess, K. R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine 13,* , 1045–1062.

Hilsenbeck, S. G., et.al. (1998). Time-dependence of hazard ratios for prognostic factors in primary breast cancer. *Breast Cancer Research and Treatment 52,*, 227–237.

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

Long J, , & Ryoo J. (2010). Using fractional polynomials to model non- linear trends in longitudinal data. *Br J Math Stat Psychol 2010; 63:* , 177-203.

Martinussen, T., , T. H. Scheike, , & I. M. Skovgaar. (2002). Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scandinavian Journal of Statistics 29(1),*, 57–74.

Ng'Andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of cox's model. *Statistics in Medicine 16,* , 611–626.

P. M. Grambsch, & T. M. Therneau. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika, 81*, 515-526, 1994.

Royston, P., & Altman, D. (1994). Regression using Fractional Polynomial of continuous covariates: Parsimonious parametric modeling. *Applied Statistics 43,*, pp. 429-467.

Schemper, M., S. Wakounig, & G. Heinze . (2009). The estimation of average hazard ratios by weighted cox regression. . *Statistics in Medicine 28,* , 2473–2489.

Schonfeld, D. . (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika 67,*, 145-53.

Schumaker, L. (1981). Spline Functions. *John Wiley, New York.*

Stone, C. (1997). Additive regression and other non-parametric models. . *Ann. Statist.* , *13*, , 689-705.

Therneau, T. M. , & P. M. Grambsch. (2000). Modeling Survival Data: Extending the Cox Model. *New York: Springer-Verlag Inc.*

Verweij, P. J., & Houwelingen, H. C. (1995). Time-dependent effects of fixed covariates in cox regression. *Biometrics 51*,, 1550–1556.

Wahba, G. (1990). Spline Models for Observational Data. *SIAM, Philadelphia*.

Yuedong. W. (2011). *Smoothing Splines Methods and Applications.* New York: Monographs on Statistics and Applied Probability 121.

# APPENDIX

## Stata codes

```
estat phtest, detail

graph combine stphplot stcoxkm

graph combine stphplot stcoxkm

graph combine stphplot stcoxkm

graph combine stphplot stcoxkm

graph combine stphplot stcoxkm

qui stcoxkm, by(gradd1) legend(cols(1)) title(Kaplan-Meier Survival Estimate) caption(Fig. 2: stcoxkm for gradd1) name(stcoxkm, replace)

qui stcoxkm, by(gradd2) legend(cols(1)) title(Kaplan-Meier Survival Estimate) caption(Fig. 2: stcoxkm for gradd2) name(stcoxkm, replace)

qui stcoxkm, by(grade) legend(cols(1)) title(Kaplan-Meier Survival Estimate) caption(Fig. 2: stcoxkm for grade) name(stcoxkm, replace)

qui stcoxkm, by(hormon) legend(cols(1)) title(Kaplan-Meier Survival Estimate) caption(Fig. 2: stcoxkm for hormon) name(stcoxkm, replace)

qui stcoxkm, by(meno) legend(cols(1)) title(Kaplan-Meier Survival Estimate) caption(Fig. 2: stcoxkm for meno) name(stcoxkm, replace)

qui stphplot, by(gradd1) title(Kaplan-Meier Survival Estimate) caption(Fig. 1: stphplot for gradd1) name(stphplot, replace)

qui stphplot, by(gradd2) title(Kaplan-Meier Survival Estimate) caption(Fig. 1: stphplot for gradd2) name(stphplot, replace)

qui stphplot, by(grade) title(Kaplan-Meier Survival Estimate) caption(Fig. 1: stphplot for grade) name(stphplot, replace)

qui stphplot, by(hormon) title(Kaplan-Meier Survival Estimate) caption(Fig. 1: stphplot for hormon) name(stphplot, replace)
```

```
qui stphplot, by(meno) title(Kaplan-Meier Survival Estimate) caption(Fig. 1: stphplot for meno) name(stphplot, replace)
stcox age meno size grade gradd1 gradd2 nodes enodes pgr er hormon
stcox age meno size grade gradd1 gradd2 nodes enodes pgr er hormon, nolog
stcoxkm, by(meno) legend(cols(1))
stcoxkm, by(meno) legend(cols(1)) title(Kaplan-Meier Survival Estimate)
stcoxkm, by(meno) legend(cols(1)) title(Kaplan-Meier Survival Estimate) caption(Fig. 2: stcoxkm for meno)
stphplot, by(meno)
stphplot, by(meno)
stphplot, by(meno) title(Kaplan-Meier Survival Estimate)
stphplot, by(meno) title(Kaplan-Meier Survival Estimate) caption(Fig. 1: stphplot for meno)
```

## Simulation 1

```
set seed 36577538
set obs 1000
gen trt=rbinomial(1,0.5)
survsim stime event, dist(weibull) cr ncr(2) lambdas(0.1 0.1) gammas(1.5 0.5) cov(trt-0.5 0.5)
replace event=0 if stime>15
stset stime, failure(event==1)
streg trt, didt(w) nohr nolog noheader
stcompet ci1=ci, compet1(2) by(trt)
stset stime, failure(event==2)
streg trt, dist(w) nohr nolog noheader
stcompet ci2=ci, compet1(1) by(trt)
```