Stockholm University

Master thesis

Department of Statistics

Masteruppsats, Statistiska institutionen

Flexible Parametric Survival Modeling with Application to the Age, Gene/Environment Susceptibility (AGES) – Reykjavik Study Data

Bergdís Björk Sigurjónsdóttir

Masteruppsats 30 högskolepoäng, vt 2015

Supervisor: Gebrenegus Ghilagaber

Abstract

This study aims to evaluate flexible parametric survival models and illustrate them using data from the Age, Gene/Environment Susceptibility (AGES) – Reykjavik Study. The flexible parametric survival models are compared with the Cox model and the standard parametric models. The goal is to examine if better insights can be gained by modeling the given data with flexible parametric models as an alternative to the standard models. The parameter estimates from the flexible modeling are comparable with those from the Cox proportional hazard modeling but have the additional characteristics of parametric models. Hence, they combine the best properties both of the standard and flexible models. Comparison and evaluation is done in such a way so that the user of the data can assess if the modifications in the data modeling process are beneficial.

Multiple models are fitted using data from about 11.537 members from the AGES study conducted by the Icelandic Heart Association. The best fit of the flexible parametric models is attained by the Weibull extension, the proportional hazard model with 3 degrees of freedom, PH(3). The results show a substantial gain in fit compared with the standard Weibull model and the parameter estimates for the hazard ratios are comparable with those from the Cox model. In tests of calibration and concordance, the PH(3) model and the Cox model fit the data equally well. It is demonstrated how the PH(3) model can be used for predictions beyond the study observation period - a feature that the Cox model lacks.

Based on the results, we recommend the use of the PH(3) model to analyze the data from the AGES study as the PH(3) model fits the data at least as good as the Cox model and additionally offers the predicting features of flexible parametric modeling.

Acknowledgement

I would like to express my gratitude to my supervisor Gebrenegus Ghilagaber for the useful comments, remarks and engagement through the learning process of this master thesis. Furthermore I would like to thank Thor Aspelund and the Icelandic Heart Association for introducing me to the topic, providing the data analyzed and for the great support on the way. I would like to thank my loved ones for their support throughout the entire process. Above all I would like to thank my partner, Sigurður Ágúst Einarsson, for his love and constant support.

Table of Contents

1 Introduction

The most widely used models in survival analysis during the last decades have been the semi-parametric Cox model and the standard parametric models. Many have chosen the well-known Cox model over the standard parametric models due to lack of fit those models often show. However, there are disadvantages of using the Cox model. It works well in estimating the hazard or survival of one group compared with another if the hazard is proportional but it leaves out estimation of the baseline hazard (Royston & Lambert, 2011). Hjort (1992) pointed out that the success of the Cox model may have had the unintended side-effect that the baseline hazard is too rarely studied. He pointed out that a parametric version of the Cox model would lead to more precise estimation of survival probabilities and it would then additionally contribute to a better understanding of the phenomenon under study.

In 2002 Royston and Parmar published an article and introduced new models for survival analysis, flexible parametric proportional-hazards and proportional-odds models. These new models are extensions of some of the standard parametric models and have been shown to give a better fit to the data than the standard ones in many situations. Furthermore the flexible proportional-hazard model is similar in parameter estimation to the Cox model and additionally estimates the baseline hazard parametrically.

The Icelandic Heart Association has been working with the data from the AGES study and the data has mainly been modeled through the use of the Cox model. Due to issues such as lack of estimation of the baseline hazard and therefore difficulties in predictions from the model, they are interested in seeing what can be gained by fitting the recent flexible parametric models.

In recent years the use of parametric survival models has been increasing in applied research. The benefits of these models have become more recognized and the availability of more flexible methods has become more accessible in standard software (Crowther & Lambert, 2014).

Royston and Parmar (2002) applied their new models to two data sets in cancer research in 2002. In this article they focused on showing that through the flexible parametric models it is simple to model the hazard function, which often has been considered problematic, and showed the benefits of it, for example in clinical trials.

In 2009 Lambert and Royston used the method of flexible parametric models in a study of breast cancer survival and incidence of hip fracture in prostate cancer patients. They compared estimates from a Cox model to estimates from a flexible parametric model which models on the proportional hazard and showed that the estimates were very similar. They also showed that there are several advantages associated with the flexible parametric approach compared with the Cox model, especially with reference to predictions.

Since 2009 flexible parametric models have been applied to several circumstances in survival analysis with good result. It has been extended to be used in relative survival, and survival analysis dealing with all cause or cause-specific survival. A flexible parametric cure model has been developed and a flexible parametric relative survival model for estimating life expectancy and loss in expectation of life has been applied (Andersson, 2013).

The purpose of this study is to evaluate the different methods of survival analysis and apply to the given data. The recent method of flexible parametric survival models is examined and compared with the Cox model and the standard parametric models. It will be shown how some of the disadvantages with the Cox model and the standard parametric models can be resolved through the use of the flexible parametric models. The goal is to estimate if a considerable improvement in fit can be gained by modeling the AGES data with flexible parametric models compared with standard parametric models. The parameter estimates should be in agreement with the proportional hazard estimates produces by the Cox model and additionally have the characteristics of parametric models. In that way it might be possible to combine the best of both models.

It is of interest to the Icelandic Heart Association to fit the flexible parametric models to the data provided, and estimate if it is a viable improvement opportunity to begin modeling its data with flexible parametric models instead of the Cox model. These types of models have not been applied to this data before and therefore it is interesting to see the fit in comparison to standard parametric models and the Cox model, which has mainly been applied to the data previously.

The research questions of this study are as follows:

- Do the flexible parametric models give a good fit to the data?
- Is the fit of the chosen flexible model better than what can be attained by a standard parametric model?
- Are the parameter estimates for the hazard ratios of the covariates in agreement with the ones accomplished by the Cox model?

The remainder of the paper is organized as follows. In section two the standard survival models, the parametric models and the Cox model, are studied and the advantages and disadvantages of the models discussed. In section three the flexible parametric models are studied as an alternative method to the standard methods and the theory behind the method explained. Measurements of goodness of fit for survival models are examined in section four. Section five covers the AGES study and the database with explanation about the variables and analysis in this paper. In section six the results are introduced and section seven follows with discussion and conclusion.

2 Standard models for survival data

2.1 Survival analysis

Survival analysis is a class of statistical methods used to analyze data in the form of times from a well-defined time origin until the occurrence of some particular event or some sort of an end point. Time origin often corresponds to the recruitment of an individual into a study. The event or end point is often considered as the death of the individual, but can also represent something else. It could for example be relief of pain or the recurrence of symptoms (Collett, 2003). Allison (2010) describes an event as a “qualitative change that can be situated in time”. By that he means a transition from one discrete state to another. When the event is death, the change will be transitioning from the state of being alive to the state of being dead. In this thesis the time origin is considered as the recruitment into the AGES study in 2002 and the event of interest is death.

An important feature of survival data which makes it difficult to handle with conventional statistical methods is censoring. When the end point of interest (in this case death) has not

been observed for an individual, the individual's survival time is said to be censored. This happens mainly for two reasons. First, some individuals may not have experienced the event of interest when the data is being analyzed, e.g. some individuals are still alive when the study is terminated. Second, individual's survival status may be unknown because that individual has been lost to follow-up. This could e.g. be an individual that was originally in the study but has then moved away and contact has been lost (Collett, 2003). All methods of survival analysis allow for censoring where procedures are applied that combine the information in the censored and uncensored cases in a way that produces consistent estimates of the parameters of interest. This can be accomplished by the method of maximum likelihood or partial likelihood (Allison, 2010).

When survival data is being described, the survival function and the hazard function are of main interest. Apart from those we have the cumulative distribution function and the probability density function. The actual survival time of an individual, *t*, is the value of a variable, *T*, which can take any non-negative value. The different values of *T* have a probability distribution and *T* is called the random variable associated with the survival time. The random variable *T* has an underlying probability density function *f(t)*. The distribution function of *T* is given by

$$F(t) = \mathrm{P}(T < t) = \int_0^t f(u)\,\mathrm{d}u.$$

F(t) represents the probability that the survival time is less than some value *t* (Collett, 2003).

The survival function of *T* is given by

$$S(t) = \mathrm{P}(T \geq t) = 1 - F(t).$$

S(t) is defined as the probability that the survival time is greater than or equal to *t*. It therefore represents the probability that an individual survives from the time origin to a time equal to or beyond *t* (Collett, 2003).

The probability density function is a common way to describe probability distributions in the case of continuous variables. It is given by

$$f(t) = \frac{\mathrm{d}F(t)}{dt} = -\frac{\mathrm{d}S(t)}{dt} \tag{2.1}$$

f(t) is the derivative, or the slope, of the cumulative distribution function, *F(t)* (Allison, 2010).

The hazard or risk of death at a time *t* is expressed through the hazard function. The hazard function of *T* is given by

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{\mathrm{P}(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}.$$

h(t) is defined as the probability that an individual dies at time *t*, conditional on survival up to that time. The cumulative hazard of *T* is given by

$$H(t) = \int_0^t h(u)\,\mathrm{d}u.$$

H(t) is featured widely in survival analysis (Collett, 2003).

The survival function, the hazard function and the probability density function are all different ways to describe the continuous probability distribution. If one is known, the other two can be derived. In equation (1) the relationship between the probability density function and the survival function is given. The hazard can also be expressed in terms of the probability density function and the survival function

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.2}$$

By combining equations (2.1) and (2.2) it can be seen that

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\{\log S(t)\}. \tag{2.3}$$

By integrating both sides of equation (2.3), an expression for the survival function in terms of the hazard function is gained.

$$S(t) = \exp\left\{-\int_0^t h(u)\mathrm{d}u\right\} = \exp\{-H(t)\}, \tag{2.4}$$

Together, equations (2.2) and (2.4) lead to

$$f(t) = h(t)\exp\left\{-\int_0^t h(u)\mathrm{d}u\right\} = h(t)\exp\{-H(t)\}$$

.

Knowing the relationships between these functions is extremely useful in survival analysis because it is often necessary to move from one interpretation to another (Allison, 2010).

2.2 Graphical methods

The first popular method to analyze survival data was the Kaplan Meier (KM) method, and today it is still normally the first step in the analysis of ungrouped censored survival data. To obtain the KM estimates of the survival function a series of time intervals is constructed such that one death time is contained in each interval. The death time is taken to occur at the start of the interval (Collett, 2003). In the absence of censored data, the KM estimator is intuitive and simple. $\hat{S}(t)$ is then simply the proportion of observations in the sample with event times greater than *t*. In the presence of censored observations things get more complicated. There are k distinct event times. At each time t_j, there are n_j individuals at risk, that have not yet experienced the event nor been censored. d_j is the number of individuals who die at time t_j. Then the KM estimator is defined as:

$$\hat{S}(t) = \prod_{j:t_j \le t}\left(1 - \frac{d_j}{n_j}\right)$$

for $t_{(1)} \le t < t_{(k)}$ (Allison, 2010). A KM plot of the survival function is a step function where the estimated survival probabilities decrease, step by step, at each death time and stay constant between adjacent death times (Collett, 2003).

When estimating the cumulative hazard function, $H(t)$, graphically here is a good nonparametric method that has good small-sample properties called the Nelson-Aalen estimator (NA)

$$\widehat{H}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j}$$

As stated before there is a known relationship between the survival and cumulative hazard functions, $S(t) = \exp\{-H(t)\}$. Therefore, it is possible to change one estimation to the other, but in fact there can be a difference in the estimation between the KM and NA. Asymptotically the two estimators are equivalent but the KM estimator is superior when estimating the survival function in small samples and the NA estimator is superior when estimating the small sample cumulative hazard function (Cleves, et al., 2002).

The hazard function is the derivative of the cumulative hazard function. The NA is a step function and can therefore not be directly differentiated which makes the matter of deriving the hazard function slightly complicated. The steps of the NA cumulative hazard can be smoothed with a kernel smoother and then the hazard function can be estimated. First the hazard contribution is estimated as

$$\Delta\widehat{H}(t_j) = \widehat{H}(t_j) - \widehat{H}(t_{j-1})$$

Then the hazard can be estimated by

$$\hat{h}(t) = b^{-1} \sum_{j=1}^{D} K_t\left(\frac{t - t_j}{b}\right) \Delta\widehat{H}(t_j)$$

for some kernel function K_t and bandwith b, summated over the D times which failure occurs (Cleves, et al., 2002).

The KM and NA methods described can be useful in the analysis of a single sample of survival data, or in the comparison of two or more groups of survival times. When a model has two or more explanatory variables, e.g. sex and age, the resulting data set is too complex for the Kaplan Meier estimates, since there is no way to test for interactions (Allison, 2010). Then two different methods have been popular. First, parametric models or so called Accelerated Failure Time Models. Second, the well-known Cox model.

2.3 Cox model

Cox introduced a model called proportional hazards model in 1972 (Cox, 1972). The model is based on the assumption of proportional hazards but assumes no particular form of probability distribution for the survival times. Therefore, the model is referred to as a semi-parametric model (Collett, 2003).

The proportional hazards model can be expressed in the form

$$h_M(t) = \psi h_F(t),$$

where $h_M(t)$ could represent the hazard of death for males at time *t* and $h_F(t)$ could represent the hazard of death for females at time *t*. This holds for any non-negative value of *t*, and where ψ is a constant. This implies that the hazard of the two groups, males and females, is proportional over time. The value of ψ is the ratio of the hazards of death at

any time for males relative to females. Therefore ψ is known as the relative hazard, or hazard ratio (Collett, 2003).

An alternative way of expressing this model is to write $h_0(t)$ for the hazard function of a comparison group, e.g. for males. The hazard for other groups, here females, is then written as relative to the comparison group, $\psi h_0(t)$. The relative hazard, ψ, cannot be negative and is often set to $\psi = \exp(\beta)$, so that $\beta = \log(\psi)$. Any value of β will then lead to a positive value of ψ. Let X be an indicator variable, which takes the value 0 for the comparison group, males, and unity for females. If x_i is the value of X for the ith individual, i=1,2,…,n, the hazard function for this individual can be written as

$$h(t|\mathbf{x}_i) = h_0(t)e^{\beta x_i}.$$

This proportional hazards model for the comparison of two treatment groups can be generalized to the situation where the hazard of death at a particular time depends on the values of the covariate vector $\boldsymbol{x}$**.** The hazard of the ith individual can then be written as

$$h(t|\mathbf{x}_i) = \psi(\boldsymbol{x}_i)h_0(t),$$

where $\psi(\boldsymbol{x}_i)$ is a function of the values of the vector $\boldsymbol{x}_i$ of explanatory variables for the ith individual. The function ψ is interpreted as the hazard at time t for an individual whose vector of explanatory variables is $\boldsymbol{x}_i$, relative to the hazard for an individual for whom $\boldsymbol{x} = 0$. As before, $\psi(\boldsymbol{x}_i)$ can be written as $\exp(\beta_x \boldsymbol{x}_i)$. Then the general proportional model becomes

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\beta_x \boldsymbol{x}_i)$$

This can be re-expressed as

$$\log\left(\frac{h(t|\boldsymbol{x}_i)}{h_0(t)}\right) = \beta_x \boldsymbol{x}_i$$

Therefore the proportional hazards model can also be regarded as a linear model for the logarithm of the hazard ratio (Collett, 2003).

2.4 Parametric models

Parametric survival models assume an underlying distribution of the survival times. The models can be written in the log-time metric, known as accelerated failure-time (AFT) metric, or in the hazard metric (Cleves, et al., 2002).

Proportional hazards models are written

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i \boldsymbol{\beta}_x)$$

The distinction of this model from the Cox model is that in the parametric PH model a functional form for $h_0(t)$ is specified and not left unparameterized as in the Cox model. The baseline hazard, $h_0(t)$, can follow Weibull distribution for example. These models produce results that are directly comparable to results by Cox regression and this comparability is probably the most attractive characteristic of the parametric PH model. Additionally it produces estimates of the ancillary parameters for the assumed distribution and from that the predicted baseline, $h_0(t)$, the baseline hazard function, can be obtained (Cleves, et al., 2002).

The accelerated failure-time models are written

$$\ln(t_i) = \mathbf{x}_i \boldsymbol{\beta}_x + \ln(T_i)$$

where T_j can follow different distributions. The word accelerated is used because a distribution is assumed for $T_i = \exp(-\mathbf{x}_i\boldsymbol{\beta}_x)\, t_i$ and $\exp(-\mathbf{x}_i\boldsymbol{\beta}_x)$ is called the acceleration parameter. If $\exp(-\mathbf{x}_i\boldsymbol{\beta}_x) = 1$ then $T_i = t_i$ and time passes at its normal rate. If $\exp(-\mathbf{x}_i\boldsymbol{\beta}_x) > 1$ then $T_i < t_i$ and time passes more quickly for the subject and failure is expected to occur sooner. Finally, if $\exp(-\mathbf{x}_i\boldsymbol{\beta}_x) < 1$ then $T_i > t_i$ and failure is expected to occur later. $\ln(T_i)$ is a random quantity with a distribution determined by what is assumed about the distribution of T_i and it is the distribution of T_i that is specified (Cleves, et al., 2002).

Some AFT models, e.g. the Weibull, have both a hazard interpretation and an AFT interpretation but other AFT models, e.g. the loglogistic and lognormal have no natural PH interpretation.

The three parametric models which have been extended by flexible parametric models will be reviewed here, that is the Weibull model, the loglogistic model and the lognormal model.

In the PH metric the Weibull model assumes the baseline hazard of the form $h_0(t) = pt^{p-1}\exp(\beta_0)$. p is an ancillary shape parameter estimated from the data and $\exp(\beta_0)$ is the scale parameter. Under the PH model, given $\mathbf{x}_i$, a set of covariates

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i\boldsymbol{\beta}_x) = pt^{p-1}\exp(\beta_0 + \mathbf{x}_i\boldsymbol{\beta}_x)$$

This yields

$$H(t|\mathbf{x}_i) = \exp(\beta_0 + \mathbf{x}_i\boldsymbol{\beta}_x)t^p$$

$$S(t|\mathbf{x}_i) = \exp\{-\exp(\beta_0 + \mathbf{x}_i\boldsymbol{\beta}_x)t^p\}$$

Depending on the estimated parameter p, the Weibull distribution provides a variety of monotonically decreasing or increasing shapes of the hazard function. When $p = 1$ the hazard is constant, when $p < 1$ it is monotone decreasing and monotone increasing when $p > 1$. Hence, when modeling data that exhibits monotone hazard rate, the Weibull distribution can be a suitable choice (Cleves, et al., 2002).

In the AFT metric the loglogistic model assumes that T_i is distributed as loglogistic with parameters (β_0, γ). Thus

$$\ln(t_i) = \mathbf{x}_i\boldsymbol{\beta}_x + \ln(T_i) = \beta_0 + \mathbf{x}_i\boldsymbol{\beta}_x + u_i$$

where u_i follows a logistic distribution. By accelerating the effect of time on survival experience the AFT formulation can be derived. With $\boldsymbol{x} = 0$, the baseline survival function of t_i is

$$S_o(t_i) = \left[1 + \{\exp(-\beta_0)t_i\}^{\frac{1}{\gamma}}\right]^{-1}$$

In the presence of non-zero covariates the time is accelerated by a factor of $\exp(-\mathbf{x}_i\boldsymbol{\beta}_x)$

Thus

$$S(t_i|\mathbf{x}_i) = S_0\{\exp(-\mathbf{x}_i\boldsymbol{\beta}_x)t_i\}$$

$$= \left[1 + \{\exp(-\beta_0)\exp(-\mathbf{x}_i\boldsymbol{\beta}_x)t_i\}^{\frac{1}{\gamma}}\right]^{-1}$$

$$= \left[1 + \{\exp(-\beta_0 - \mathbf{x}_i\boldsymbol{\beta}_x)t_i\}^{\frac{1}{\gamma}}\right]^{-1}$$

If $\gamma < 1$ the logistic hazard is unimodal, first the hazard increases and then decreases but if $\gamma \geq 1$ the hazard is monotone decreasing (Cleves, et al., 2002).

In the AFT metric the lognormal model assumes that T_i is distributed as lognormal with parameters (β_0, σ). Thus

$$\ln(t_i) = \mathbf{x}_i\boldsymbol{\beta}_x + \ln(T_i) = \beta_0 + \mathbf{x}_i\boldsymbol{\beta}_x + u_i$$

where u_i follows a standard normal distribution. By accelerating the effect of time on survival experience the AFT formulation can be derived. With $\boldsymbol{x} = 0$, the baseline survival function of t_i is

$$S_o(t_i) = 1 - \Phi\left(\frac{\ln t_i - \beta_0}{\sigma}\right) = \Phi\left(-\frac{\ln t_i - \beta_0}{\sigma}\right)$$

where Φ is the cumulative standard normal distribution. The same way as with the loglogistic model the effect of the covariates is to accelerate time by a factor of $\exp(-\mathbf{x}_i\boldsymbol{\beta}_x)$. Thus

$$S(t_i|\mathbf{x}_i) = S_0\{\exp(-\mathbf{x}_i\boldsymbol{\beta}_x)t_i\}$$

$$= \Phi\left[-\frac{\ln\{\exp(-\mathbf{x}_i\boldsymbol{\beta}_x)t_i\} - \beta_0}{\sigma}\right]$$

$$= \Phi\left[-\frac{\ln t_i - (\beta_0 + \mathbf{x}_i\boldsymbol{\beta}_x)}{\sigma}\right]$$

The hazard function of the lognormal model is nonmonotonic, it increases and then decreases in a unimodal way (Cleves, et al., 2002).

The loglogistic and lognormal models are similar and neither has a natural PH interpretation, but the loglogistic model has a proportional odds (PO) interpretation. One advantage of the loglogistic model over the lognormal model is that the mathematical expressions of the hazard and survival functions are simpler and the expressions do not include the normal cumulative distribution function (Cleves, et al., 2002).

2.5 Drawbacks of the standard models

Both the Cox model and the standard parametric models have some weaknesses that have led to the development of more flexible models.

Royston and Parmar (2002) pointed out three issues that they have regarding the Cox model. First, that the behaviour of the hazard function is of potential interest, for instance related to the time-course of an illness. Second, they say that there is an issue with the Cox model regarding how to deal with non-proportional hazards which may occur. Although the model may be extended to allow for non-proportional hazards there is no widely accepted approach. Third, an issue arises in the validation of a survival model. Since the baseline hazard is not modeled the fitted model is too closely adapted to the

data at hand. They say that the strength of the Cox model is to be able to fit a model and get regression coefficients without thinking about the underlying distribution, but this is also the weakness of the model.

Hjort (1992) pointed out that the success of the model has perhaps had the unintended side-effect that practitioners too seldomly invest efforts in studying the baseline hazard. He said that if there would be an adequate parametric version of the Cox model it would lead to more precise estimation of survival probabilites and concurrently contribute to a better understanding of the phenomenon under study. Cox himself said in a radio interview in 1994 that he would normally wish to tackle a problem parametrically, because operations such as predictions were so much easier (see Royston and Lambert, 2011).

The standard parametric models do estimate the baseline hazard parametrically and the advantage of using the parametric approach is the ease of obtaining predictions (Lambert & Royston, 2009). The problem with many of those models is that they make strong assumptions about the shape of the baseline hazard function. The Weibull model, as previously stated, assumes a monotonically increasing or decreasing baseline hazard and the lognormal model assumes a unimodal hazard. Real data often exhibit turning points in the underlying hazard function and therefore the parametric models are often not flexible enough (Crowther & Lambert, 2014). For this reason, standard parametric models may have some theoretical advantages but they are generally not sufficiently flexible to represent real data adequately (Royston & Lambert, 2011).

An example of what can be gained with a suitable parametric model is predicting life expectancy. National statistical agencies produce national life table and estimate statistics on period life expectancy by age and sex. These are the average number of additional years a person would live if he or she would experience the age-specific mortality rates of the given time period and area for the rest of their life. However death rates are not static over time and therefore the period life expectancy is not the number of years someone in the area in that time period is actually likely to live (Office for National Statistics, 2014). The life expectancy is therefore estimated with a Kaplan-Meier type approach. Life expectancy could be predicted with a parametric survival model. Unlike the Cox model, extrapolation of survival estimates beyond the study observation period is possible using a parametric approach. Jackson et al. (2011) have shown that this can be a useful option when predicting life expectancy for people with cystic fibrosis disease. A well-fitted parametric model could therefore be used to predict life expectancy in a more accurate way.

3 Flexible parametric models

Problems with the common poor fit of standard parametric survival models and the lack of baseline hazard estimation in the Cox model have led to a new development. In 2002 Royston and Parmar published an article with new developments on more flexible parametric models. In the article they showed a way to extend commonly known parametric survival models so that the models would fit better than the standard models and still be able to model the baseline hazard. The flexible parametric models are extensions of some of the standard parametric models and are fitted through the use of restricted cubic splines.

3.1 Restricted cubic splines

Cubic splines are piecewise cubic polynomials in the form of $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ with a separate cubic polynomial fit in predefined number of intervals. The split points of the intervals are known as knots. Three restrictions are necessary so that the fitted function will be smooth. The cubic function is forced to join at the knot locations, making the function continuous. The first derivative of the spline functions are also forced to agree at the knots. Since the first derivative is the gradient of the function it should lead to smoother function to force the agreement at the knots. The last restriction is to force the second derivative to agree at the knots and since that represents the rate of change in the gradient the function becomes even smoother. Cubic splines are the most common type of splines used in practice and higher degree polynomials are generally not needed. If there is a complicated shape between knots then further knots should be added rather than fitting a higher degree polynomial (Royston & Lambert, 2011).

Let K represent number of knots at $k_1 < \cdots < k_K$ for a nonlinear spline function, $s(x)$, for covariate x. Then

$$s(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{K} \beta_{i3} (x - k_i)_+^3$$

denotes a cubic spline function without continuity restriction, where

$$u_+ = \begin{cases} u & \text{if } u > 0 \\ 0 & \text{if } u \leq 0 \end{cases}$$

The method of flexible parametric survival analysis models with restricted cubic splines. That indicates that the function is forced to be linear before the first knot and after the last knot. The first and last knots are defined as the minimum and maximum of the uncensored survival times and are known as boundary knots. In order to fit a restricted cubic spline function for a covariate x, transformations of x are included as new variables in the linear predictor. Let $s(x)$ be the restricted cubic spline function with m interior knots, $k_1, \dots, k_m$ in addition to the two boundary knots, k_{min} and k_{max}. Then $s(x)$ can be written as a function of parameters γ and the newly created variables $z_1, \dots, z_{m+1}$,

$$s(x) = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \gamma_{m+1} z_{m+1}$$

The z_j variables are calculated as follows

$$z_1 = x$$

$$z_j = \left(x - k_j\right)_+^3 - \lambda_j (x - k_{min})_+^3 - \left(1 - \lambda_j\right)(x - k_{max})_+^3$$

where for $j = 2, \dots, m + 1$,

$$\lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}}$$

The linearity restriction goes beyond the observed data because the boundary knots are the minimum and maximum event times. In the tails of the distribution, where data is often sparse, restricted cubic splines tend to give more believable estimates. The function of restricted cubic splines is programmed into different software, including the various Stata commands (Royston & Lambert, 2011).

3.2 The different flexible parametric models

In 2002 Royston and Parmar introduced the flexible parametric proportional-hazards (PH) and proportional-odds (PO) models. The two models are extensions of the Weibull and loglogistic parametric models. Royston and Parmar also mentioned the flexible probit-scale model, an extension of the lognormal parametric model. These alternatives extend the range of survival distributions that can be estimated and today the PH, PO and probit-scale models are still the only flexible parametric models that have interpretable covariate effects. It is assumed that the effect of covariates is proportional on the appropriate scale, hazard, odds of failure or probit of failure probability, implying linearity between certain transformation of the survival function and the logarithm of survival time. In the flexible parametric survival models the linearity restriction of the transformed survival function ($ln(t)$) is relaxed and nonlinear functions allowed (Royston & Lambert, 2011).

The basic Weibull model with covariate vector **x** and parameter vector $\boldsymbol{\beta}$ can be written as

$$\ln H(t|\mathbf{x}) = \ln H_0(t) + \mathbf{x}\boldsymbol{\beta} = \gamma_0 + \gamma_1 \ln t + \mathbf{x}\boldsymbol{\beta}$$

The baseline cumulative hazard function can be extended to a restricted cubic spline function of $\ln t$, $\ln H_0(t) = s(\ln t; \gamma)$. The extension of the Weibull model to a more general PH model is then

$$\ln H(t|\mathbf{x}) = \ln H_0(t) + \mathbf{x}\boldsymbol{\beta} = s(\ln t|\gamma) + \mathbf{x}\boldsymbol{\beta}$$

where

$$s(\ln t|\gamma) = \gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + \cdots$$

and $\ln t$, $z_1(\ln t)$, $z_2(\ln t)$ and so on, are the basis functions of the restricted cubic splines.

By exponentiating and differentiating with respect to t we get the hazard function and then the log hazard function by taking the logarithm

$$h(t|\mathbf{x}) = \frac{ds(\ln t\,|\gamma)}{dt} \exp\{s(\ln t|\gamma) + \mathbf{x}\boldsymbol{\beta}\}$$

$$\ln h(t|\mathbf{x}) = \ln\left\{\frac{ds(\ln t\,|\gamma)}{dt}\right\} + s(\ln t|\gamma) + \mathbf{x}\boldsymbol{\beta}$$

By evaluating the derivative (see Appendix A) it can be seen that

$$\ln h(t|\mathbf{x}) = -\ln t + \ln(\gamma_1 + \gamma_2 z_2' + \cdots + \gamma_{m+1} z_{m+1}') + \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \gamma_{m+1} z_{m+1} + \mathbf{x}\boldsymbol{\beta}$$

The proportional-hazard model is referred to as PH(*d*) model where *d* represents the degrees of freedom. A model has *d*-1 interior knots and when *d* >1, two boundary knots. The Weibull model can therefore be referred to as PH(1), but with different parametrization (Royston & Lambert, 2011).

The odds ratio (OR) is often used as an approximate measure of the change in risk of an event, such as death, occurring in the presence of some factor of interest. In 2x2 tables OR is defined by

$$\text{OR} = \frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}$$

p_1 and p_2 here represent the proportion of individual with event of interest in different groups. Generalizing to the case where failure occurs with probabilities $F_1(t)$ and $F_2(t)$ by time $t > 0$ and rotating we get

$$\frac{F_1(t)}{1-F_1(t)} = \frac{F_2(t)}{1-F_2(t)} \text{OR}(t)$$

If $\text{OR}(t)$ is constant then the model has the assumption of proportional odds. That implies that the odds of an event occurring up to time t are proportional across levels of the covariate (Royston & Lambert, 2011).

The PO model can be generalized to the situation where individual i has a covariate vector $\mathbf{x}_i$ with parameter vector $\boldsymbol{\beta}$. Let $\text{OR}_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$, then

$$\frac{F(t|\mathbf{x}_i)}{1-F(t|\mathbf{x}_i)} = \frac{F_0(t)}{1-F_0(t)} \exp(\mathbf{x}_i\boldsymbol{\beta})$$

where $F_0(t) = F(t|\mathbf{0})$ is the baseline distribution function and $F(t|\mathbf{x}_i)$ is the cumulative distribution function at time t. Since $S(t|\mathbf{x}_i) = 1 - F(t|\mathbf{x}_i)$ and $\text{logit}(x) = \ln\{x/(1-x)\}$, the PO model can be expressed as

$$\text{logit}\{1 - S(t|\mathbf{x}_i)\} = \text{logit}\{1 - S_0(t)\} + \mathbf{x}_i\boldsymbol{\beta} \qquad (3.1)$$

The loglogistic model is a PO model with the baseline survival function

$$S_o(t) = \left[1 + \{\exp(-\beta_0)t\}^{\frac{1}{\gamma}}\right]^{-1}$$

This can be expressed as

$$\text{logit}\{1 - S_0(t)\} = (-\beta_0 + \ln t)/\gamma = \gamma_0 + \gamma_1 \ln t$$

where $\gamma_0 = -\beta_0/\gamma$ and $\gamma_1 = 1/\gamma$. The loglogistic model is an AFT model with parameter vector $\boldsymbol{\beta}^*$ defined by

$$S(t|\mathbf{x}_i) = S_0\{\exp(-\mathbf{x}_i\boldsymbol{\beta}^*)t\}$$

giving

$$\begin{aligned}\text{logit}\{1 - S(t|\mathbf{x}_i)\} &= \text{logit}[1 - S_o\{\exp(-\mathbf{x}_i\boldsymbol{\beta}^*)t\}] \\ &= \gamma_0 + \gamma_1 \ln\{t \exp(-\mathbf{x}_i\boldsymbol{\beta}^*)\} \\ &= \gamma_0 + \gamma_1 \ln t - \mathbf{x}_i\boldsymbol{\beta}^*\end{aligned}$$

Thus

$$\text{logit}\{1 - S(t; \mathbf{x}_i)\} = \text{logit}\{1 - S_0(t)\} - \mathbf{x}_i\boldsymbol{\beta}^* \qquad (3.2)$$

By writing $\boldsymbol{\beta} = -\boldsymbol{\beta}^*$ it can be seen that equations 3.1 and 3.2 are identical and therefore the loglogistic model is a PO model with different parametrization. The loglogistic model can be extended to increase flexibility in a similar way as the Weibull model by using spline functions of $\ln t$.

$$\text{logit}\{1 - S(t|\mathbf{x})\} = \text{logit}\{1 - S_0(t)\} + \mathbf{x}\boldsymbol{\beta} = \boldsymbol{s}(\ln t|\gamma) + \mathbf{x}\boldsymbol{\beta}$$

The logit of the baseline distribution function is modeled as a restricted cubic spline in $\ln t$. A proportional odds model is referred to as PO(*d*) where *d* represents the degrees of freedom. A model has *d*-1 interior knots and when *d* >1, two boundary knots. The loglogistic model can therefore be referred to as PO(1), but with different parametrization (Royston & Lambert, 2011).

The lognormal survival model can be written as

$$\ln t = \beta_0 + \mathbf{x}\boldsymbol{\beta}^* + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$ is a normally distributed residual and $\ln t \sim N(\beta_0 + \mathbf{x}\boldsymbol{\beta}^*, \sigma^2)$. The cumulative distribution and survival functions of $\ln t$ are therefore

$$F(\ln t) = \Phi\left(\frac{\ln t - \beta_0 - \mathbf{x}\boldsymbol{\beta}^*}{\sigma}\right)$$

$$S(\ln t) = 1 - F(\ln t) = \Phi\left(-\frac{\ln t - \beta_0 - \mathbf{x}\boldsymbol{\beta}^*}{\sigma}\right)$$

By writing $\boldsymbol{\beta} = -\boldsymbol{\beta}^*/\sigma$ and rotating we get

$$-\Phi^{-1}\{S(\ln t)\} = \frac{\ln t - \beta_0}{\sigma} + \mathbf{x}\boldsymbol{\beta}$$

When $\mathbf{x} = \mathbf{0}$ we get

$$-\Phi^{-1}\{S_0(\ln t)\} = \frac{\ln t - \beta_0}{\sigma}$$

Thus

$$-\Phi^{-1}\{S(\ln t)\} = -\Phi^{-1}\{S_0(\ln t)\} + \mathbf{x}\boldsymbol{\beta} \qquad (3.3)$$

A model satisfying (3.3) can be described as a probit model on the survival-probability scale. Probit is the inverse cumulative standard normal distribution function. Let $\gamma_0 = -\beta_0/\sigma$ and $\gamma_1 = 1/\sigma$ then (2.7) becomes

$$-\Phi^{-1}\{S(\ln t)\} = \gamma_0 + \gamma_1 \ln t + \mathbf{x}\boldsymbol{\beta}$$

This can be generalized to

$$-\Phi^{-1}\{S(\ln t)\} = \mathbf{s}(\ln t\,|\boldsymbol{\gamma}) + \mathbf{x}\boldsymbol{\beta}$$

This models the probit of the baseline survival distribution as restricted cubic splines and is denoted by probit(*d*) with *d* > 1 degrees of freedom (Royston & Lambert, 2011).

3.3 Number and position of knots

Restricted cubic splines with m interior knots and 2 boundary knots have been introduced. As stated earlier the two boundary knots are chosen to be the smallest and largest uncensored log survival-times. The next issue is how to choose the position and number of the interior knots. According to Royston and Lambert (2011) optimal knot positioning does not appear to be critical for a good fit and can even be considered undesirable. The fitted curve may then follow small-scale features of the data too closely and knots will become additional model parameters if selected data-driven. Royston and Parmar (2002) suggested positioning the knots on the centiles of the distribution of uncensored log event-times. If the model has one interior knot it should be placed on the 50^{th} centile, if there are three interior knots they should be placed on the 25^{th}, 50^{th} and 75^{th} centile, and so on. According to Royston and Lambert (2011) a worthwhile improvement in fit can be gained by using a spline model with a single interior knot (2 d.f.) over a model with no knot (the standard parametric models). Often little is gained by adding further knots. They recommend spline models with 2 or 3 degrees of freedom as a reasonable initial or default choice for smaller datasets. For larger datasets they suggest looking informally at the AIC or BIC of models with between 1 and 6 d.f. AIC is defined to be the deviance plus a penalty of twice the number of model parameters, k. BIC has a more stringent criterion, a penalty of $\ln n$ times the number of model parameters, where n is the number of events.

$$AIC = -2 \ln L + 2k$$

$$BIC = -2 \ln L + k \ln n$$

The preferred model should be the one that minimizes AIC or BIC but this criterion should not be applied mechanically in the interest of parsimony and to reduce overfitting. The aim is not necessarily to find an optimal fit but rather to capture the behavior of the data. Note that the AIC and BIC cannot be compared between a (flexible) parametric model and the Cox model. The former is fit by maximum likelihood while the latter is fit by maximum partial likelihood. Therefore are the values of AIC or BIC not comparable (Royston & Lambert, 2011).

3.4 Parameter estimation through the likelihood function

As already stated, there is a difference between the flexible parametric models and the Cox model how the parameters are estimated. The Cox model estimates the parameters through partial maximum likelihood while the flexible parametric models estimate through full maximum likelihood like the standard parametric models do.

For the flexible parametric models we define δ_i as 1 for an observed event and 0 for a right-censored observation. The sample comprises n independent observations $\{t_i, \delta_i, \boldsymbol{x}_i\}$ and the likelihood for the ith observation is l_i. The likelihood for the whole sample is then $\prod_{i=1}^{n} l_i$. The contribution of the ith observation to the total log likelihood is $\delta_i \ln h(t_i) + \ln S(t_i) - \ln S(t_{0i})$ where t_{0i} represents a possible late entry. Let $\eta_i = s(\ln t_i|\boldsymbol{\gamma}) + \boldsymbol{x}_i\boldsymbol{\beta}$ and the first derivative

$$\eta_i' = ds(\ln t_i|\boldsymbol{\gamma})/dt_i = t_i^{-1}\, ds(\ln t_i|\boldsymbol{\gamma})/d(\ln t_i).$$

Also,

$$\frac{ds(\ln t_i|\boldsymbol{\gamma})}{d(\ln t_i)} = \gamma_1 + \sum_{j=2}^{m} \gamma_j \frac{dz_j(\ln t_i)}{d(\ln t_i)}$$

$$= \gamma_1 + \sum_{j=2}^{m} \gamma_j \left\{ 3\left(\ln t_i - k_j\right)_+^2 - 3\lambda_j (\ln t_i - k_{min})_+^2 - 3\left(1-\lambda_j\right)(\ln t_i - k_{max})_+^2 \right\}.$$

For PH models the likelihood becomes

$$l_i = \begin{cases} \eta_i' \exp(\eta_i - \exp \eta_i) & \text{for an observed event,} \\ \exp(-\exp \eta_i) & \text{for a censored observation.} \end{cases}$$

The expression for the observed event is the density function at t_i and the estimated survival probability at t_i for the censored observation.

The parameters are estimated by maximum likelihood. Suitable starting values, to obtain estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, are acquired by fitting a Cox model with the covariates **x**. The survival function $\hat{S}(t_i|\boldsymbol{x}_i)$ is estimated from the Cox model as the baseline survival function at t_i raised to the power of the estimated relative hazard, $\exp\left(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}}\right)$. In the PH model the initial guess is $\ln\left\{-\ln \hat{S}(t_i|\boldsymbol{x}_i)\right\}$ for $\ln H(t_i|\boldsymbol{x}_i)$. The starting values of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are then determined by ordinary least-squares regression of these functions on $\ln t_i, \boldsymbol{x}_i$ and the spline basis functions with the desired number of knots. Full maximum likelihood estimation is then performed (Royston & Lambert, 2011).

3.5 Goodness of fit

Measurements of explained variation, R^2, are common in many fields of statistics. Measurements of such are not so straight forward in the presence of censored observation as is common with survival analysis data. Several authors have proposed versions of explained variation statistic for use with survival data (Royston & Lambert, 2011).

Royston and Sauerbrei (2004) introduced the D statistic as a measure of variation in outcome among individuals on the appropriate scale, where the D stands for discrimination. The estimated prognostic index, $\boldsymbol{x}\widehat{\boldsymbol{\beta}}$, is ordered as well as the rankits, the expected normal order statistics corresponding to these values calculated. The rankits are then scaled by a factor $\kappa = \sqrt{8/\pi}$ and an auxiliary regression performed on the scaled rankits. This ensures that D has the character of a log hazard ratio between equal-sized prognostic groups. The estimated regression coefficient is then D. For D to be useful, the prognostic index must be approximately normally distributed. This is usually fulfilled by the central limit theorem.

Royston and Sauerbrei (2004) then introduce R_D^2 as a transformation of the D statistic.

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

and

$$\sigma^2 = \begin{cases} 1 & \text{for models with a probit link} \\ \pi^2/3 & \text{for PO models} \\ \pi^2/6 & \text{for PH models and Cox model.} \end{cases}$$

The R_D^2 and D statistics are by construction only sensitive to the ranks of $\boldsymbol{x}\widehat{\boldsymbol{\beta}}$ and not its actual values. A minor change in the model that leaves the rank order unchanged does therefor not affect R_D^2 or D. This leads to robustness to outliers but the disadvantage is that the auxiliary model on the rankits may not fit the data perfectly. This might result in underestimation of R_D^2 and D.

An alternative measure of discrimination is Harrell's concordance statistic. It measures the agreement of predictions with observed failure order and is defined as the proportion of all usable subject pairs in which the predictions and outcomes are concordant, that is, predict the same outcome as occurs (Cleves, et al., 2002; Royston & Lambert, 2011).

Measurement of calibrations is another common model performance characteristic of prediction models. The Hosmer-Lemeshow (HL) goodness-of-fit test was developed to estimate the calibration of logistic regression models (Hosmer & Lemeshow, 1980). Nam and D'Agostino extended the HL test for survival data but Demler, Paynter and Cook (2015) have shown that the Nam-D'Agostino (ND) test is sensitive for censoring and the more censoring there is, the less stable the test statistic becomes. Demler et al. extended the ND test statistic by using a different variance estimator. In the original ND test a binary proportion estimator was used as a measurement of the variance and that does not account for censoring or for time of event. Demler et al. proposed using the Greenwood formula for the variance of failure probability instead. The Greenwood variance estimator of the Kaplan-Meier failure probability in the gth decile of risk scores at time t, $KM_g(t)$ is:

$$Var\left(KM_g(t)\right) = KM_g(t)^2 \sum_{i|t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

Where, as previously stated, d_i and n_i represent the number of events and number at risk at time t_i. The test statistic proposed by Demler et al., the Greenwood-Nam-D'Agostino (GND) is:

$$\chi^2_{GND}(t) = \sum_{j=1}^{G} \frac{\left[KM_g(t) - \overline{p(t)_g}\right]^2}{Var\left(KM_g(t)\right)} \sim \chi^2_{G-1}$$

The term $\left(\overline{p(t)_g}\right)$ represents the mean predicted probability of failure for subjects in the gth decile. The numerator therefore represents the squared difference of observed and expected failures. In a simulation study Demler et al. tested the performance of the GND test for variety of situations, with censoring rates of 25% and 50%, event rates of 0,05, 0,1 and 0,4 and for decreasing, increasing and constant baseline hazard. In all these situations the performance of the test was good.

4 Application to the AGES Data

4.1 The AGES study

In 1967 a cohort was established for a study of cardiovascular disease in Iceland, The Reykjavik study. The cohort comprised all men and women in the Reykjavik area born in 1907-1935. The cohort was randomly divided into six groups and the sample examined comprised all groups except one, total of 26.489 individuals. In 2002 11.549 previously examined Reykjavik study cohort members were still alive. A random sample of 8.000 surviving participants of the Reykjavik study was invited to participate in The Age Gene/Environment Susceptibility - Reykjavik Study (AGES) between 2002 and 2006. 5.764 participated (Harris, et al., 2007). Information on the variables used in the current study is available for all the 11.549 surviving participants through the registries of Statistics Iceland and will therefore be used.

4.2 Variables

The variables used in this study are the following ones:

- Agebegin02: Age in the year 2002 measured as integer.
- Lifetime: Age at death.
- Fu02y: Length of follow-up. Calculated as the difference between lifetime and agebegin02. Follow-up ended in the end of 2013.
- Death: Indicator variable where 1 represents event and 0 represents censoring.
- Sex: Sex of the participants. 1 represents males and 2 represents females.

Tables 4.1 and 4.2 show the descriptive statistics for the variables in the dataset.

Table 4.1. Descriptive statistics of the continuous variables.

	Age in 2002	Lifetime	Follow-up
Min	66	66,19	0,0082
Max	96	103,82	11,80
Mean	75,97	84,59	8,62
Sd	6,53	5,71	3,85
Median	75	84,31	10,96

Table 4.2. Descriptive statistics of the categorical variables.

Variable	Category	n	Proportion
Sex	Male	4791	0,42
	Female	6746	0,58
Death	Event	6180	0,54
	Censored	5357	0,46

4.3 Data analysis

When investigating the data on the 11.549 participants it was observed that 12 participants had lifetimes that were equal or shorter than their age in the year of 2002. According to that they should not have been alive by the beginning of the study. The reason for this is that at a certain time interval in the 20th century it happened that personal numbers of diseased individuals were reused if needed. Therefore is the information on these 12 individuals not correct. These individuals were therefore removed from the dataset, which then consists of 11.537 individuals.

Most data analysis in this study is performed with the data analysis and statistical software Stata 12. Currently it is the only statistical software that can perform an analysis of flexible parametric survival models. The model applied in this study will include the age in the year of 2002 and sex as explanatory variables and the length of follow-up as dependent variable. A SAS macro is used to test the calibration of the model (Demler, et al., 2015).

The critical p-value used in the significance tests of the data is 0,05.

5 Results

In Table 5.1 the age variable has been divided into six age groups, all with a range of five years per group except for the first group which ranges from 66-71 or six years.

Table 5.1. Event and deaths per 1.000 person-years per sex and age group.

Sex	Age	Death	Censored	Proportion Death	Proportion Censored	Person-years of follow-up	Deaths per 1.000 person-years
Male	66-71	471	997	0,32	0,68	14875,51	31,66
	72-76	658	649	0,50	0,50	11962,02	55,01
	77-81	821	291	0,74	0,26	8144,68	100,80
	82-86	559	48	0,92	0,08	3361,66	166,29
	87-91	248	4	0,98	0,02	980,98	252,81
	92-96	45	0	1,00	0,00	125,48	358,63
Female	66-71	478	1521	0,24	0,76	21330,28	22,41
	72-76	645	1085	0,37	0,63	17145,17	37,62
	77-81	919	576	0,61	0,39	12605,50	72,90
	82-86	796	163	0,83	0,17	6327,58	125,80
	87-91	447	23	0,95	0,05	2167,33	206,24
	92-96	93	0	1,00	0,00	346,39	268,48
Total		6180	5357	0,54	0,46	99372,58	62,19

As can be seen there is a considerable difference in proportions that experience the event between the age groups going from 0,32 up to 1 for males and from 0,24 up to 1 for females. There is a difference between the sexes as well in all age groups except for the last one where everyone experiences the event. The difference between the sexes ranges from 3-13 percentage points in the other age groups. Deaths per 1.000 person-years also vary between the sexes and the different age groups. Males have more deaths per 1.000 person-years in all age groups compared with females, where the group of males experiences between 23-46 percent more deaths per 1.000 person-years than the group of females. As would be expected, increasing numbers of deaths are observed per 1.000 person-years with each age group.

Figure 5.1 shows the Kaplan Meier estimates of the survival functions between different levels of the covariates. Log Rank test of the equality over the different levels of the sex variable is significant with a p-value < 0,0001 and it is also significant over the different ages with a p-value < 0,0001. When the equality is tested over the different combinations of sex and age the p-value is < 0,0001. Therefore it can be stated that there is a difference in survival between sex and age.

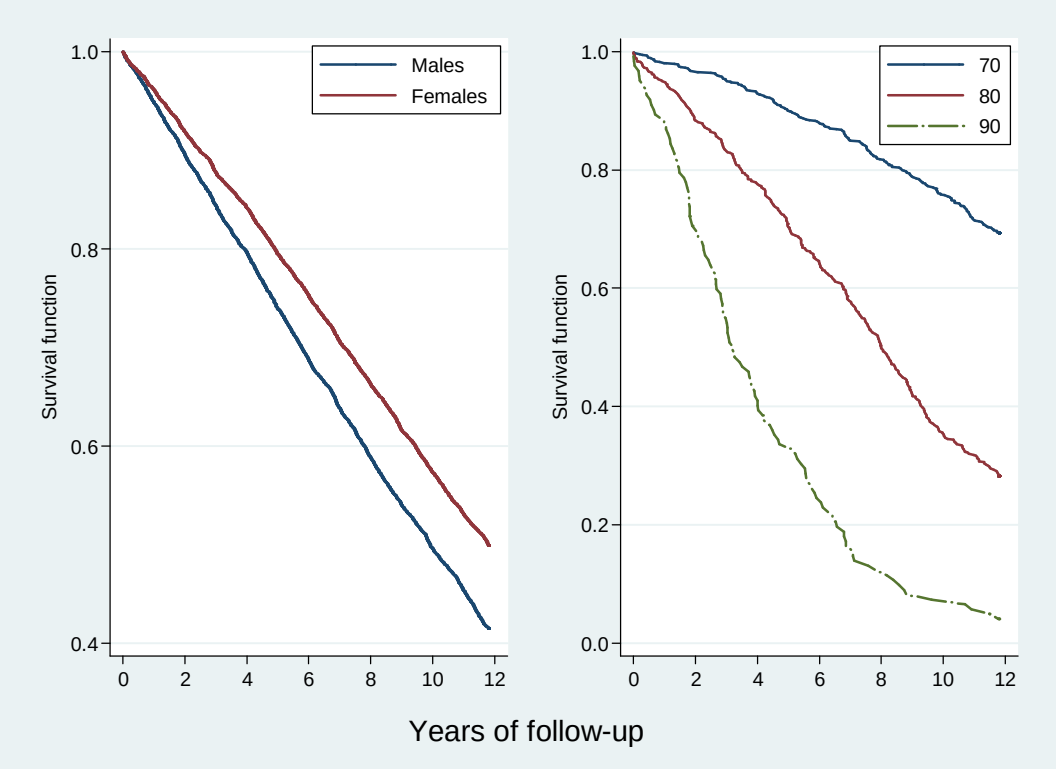

Figure 5.1. KM estimates of the survival functions of the sexes and at the age of 70, 80 and 90.

5.1 Selection of model

In Table 5.2 a comparison in AIC and BIC statistics is shown between the different flexible parametric models with 0-5 interior knots (1-6 df). As mentioned earlier PH(1) is the Weibull model, PO(1) is the loglogistic model and Probit(1) is the lognormal model. PH(4) has the lowest AIC value and PH(3) has the lowest BIC value. The difference in AIC between PH(3) and PH(4) is 0,75 but the difference in BIC is 6,6. On the grounds of parsimony the preferred model here is PH(3).

Table 5.2. AIC and BIC statistics for the different models. 24.600 is subtracted from all the values.

	PH		PO		Probit	
df	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
1	274,75	304,16	847,24	874,16	1763,31	1790,23
2	83,09	119,86	394,52	428,17	629,96	663,60
3	47,22	91,34	281,06	321,44	462,73	503,10
4	46,47	97,94	275,91	323,01	453,00	500,10
5	47,15	105,98	274,03	327,87	448,58	502,41
6	49,17	115,35	275,57	336,13	449,31	509,88

Table 5.3 contains a comparison in parameter estimates between PH models with 1-6 degrees of freedom and parameter estimates from the Cox model. As can be seen the parameter estimate for age (β_1) is the same for the Cox model and PH models with three or more degrees of freedom and the parameter estimate for sex (β_2) is almost the same for the Cox model and PH models with three or more degrees of freedom. This further strengthens the choice of the PH(3) model.

Table 5.3. Parameter estimation of age and sex in PH models with 1-6 df and the Cox model.

Model	β_1	Se(β_1)	β_2	Se(β_2)
Weibull	1,1288	0,0023	0,7133	0,0183
PH (2)	1,1323	0,0023	0,7077	0,0182
PH (3)	1,1334	0,0023	0,7064	0,0182
PH (4)	1,1334	0,0023	0,7063	0,0181
PH (5)	1,1334	0,0023	0,7063	0,0181
PH (6)	1,1334	0,0023	0,7063	0,1810
Cox	1,1334	0,0023	0,7063	0,0181

As the name indicates, one of the assumptions of PH models is that the effects of the covariates are proportional between their levels. To test if the age and sex variables violate the assumption of proportional hazard a test based on Schoenfeld residuals was conducted. The resulting p-value for the sex variable was 0,5801 and for the age variable 0,8982. Therefore, no evidence is found that the variables violate the PH assumption. In Figure 5.2 the smoothed hazard estimates of males and females can be seen. The figure reveals that the PH assumption of males and females is appropriate.

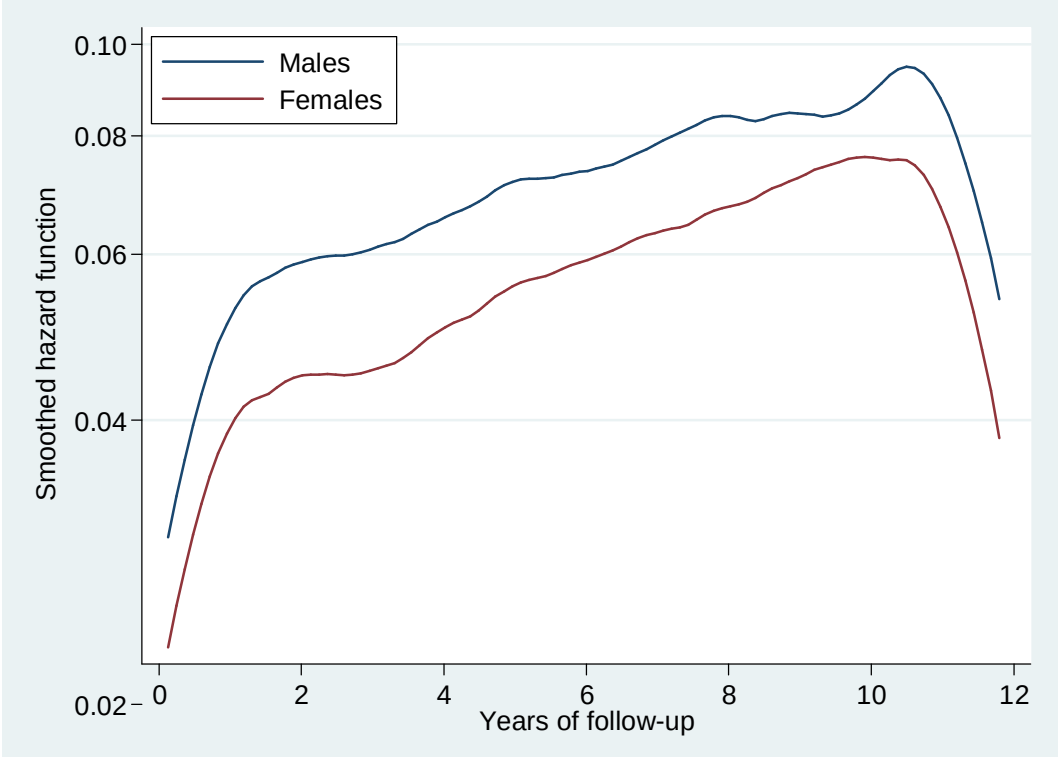

Figure 5.2. Smoothed hazard estimates for males and females.

Table 5.4 compares the explained variation and discrimination statistics between the PH models. As can be seen there is no improvement in the model by including more than two interior knots (3 df.). The variation in events during the follow-up time explained by age and sex in PH(3) is 0,313. The choice of the PH(3) model as the preferred parametric model is now apparent. Henceforth it will be the flexible parametric model used in the analysis.

Table 5.4. R^2 and discrimination statistics for PH models with 1-6 degrees of freedom.

	Degrees of freedom					
Statistic	**1**	**2**	**3**	**4**	**5**	**6**
R_D^2	0,299	0,310	0,313	0,313	0,313	0,313
D	1,338	1,372	1,382	1,382	1,382	1,382

Harrell's C is representing the proportion of all usable subject pairs in which the predictions and outcomes are concordant is 0,7131 both for the PH(3) model and the Cox model.

Table 5.5. χ^2_{GND} tests for the three moedels

Statistic	**Weibull**	**PH(3)**	**Cox**
χ^2_{GND}	17,80	12,41	12,41
P-value	0,04	0,19	0,19

The Greenwood-Nam-D'Agostino test is insignificant both for the PH(3) model and for the Cox model. That means that there is no indication that these two models lack in calibration. The Weibull model, on the other hand, tests significant, indicating differences between predicted and observed events.

5.2 Comparisons

In Figures 5.3, 5.4 and 5.5 the survival function, the cumulative hazard function and the hazard function, respectively, are compared between the Weibull model and the PH(3) model. These are plotted with what is seen in the data through the Kaplan Meier, Nelson Aalen or kernel smoothed Nelson Aalen functions, along with 95% confidence interval. All the figures show how the PH(3) model captures the behavior of the data better than the Weibull model. In Figures 5.3 and 5.4 the Weibull function lies partly outside the 95% CI and in Figure 5.5 it does not capture the right shape. The plotted functions for males and females separately indicate the same and can be found in Appendix B.

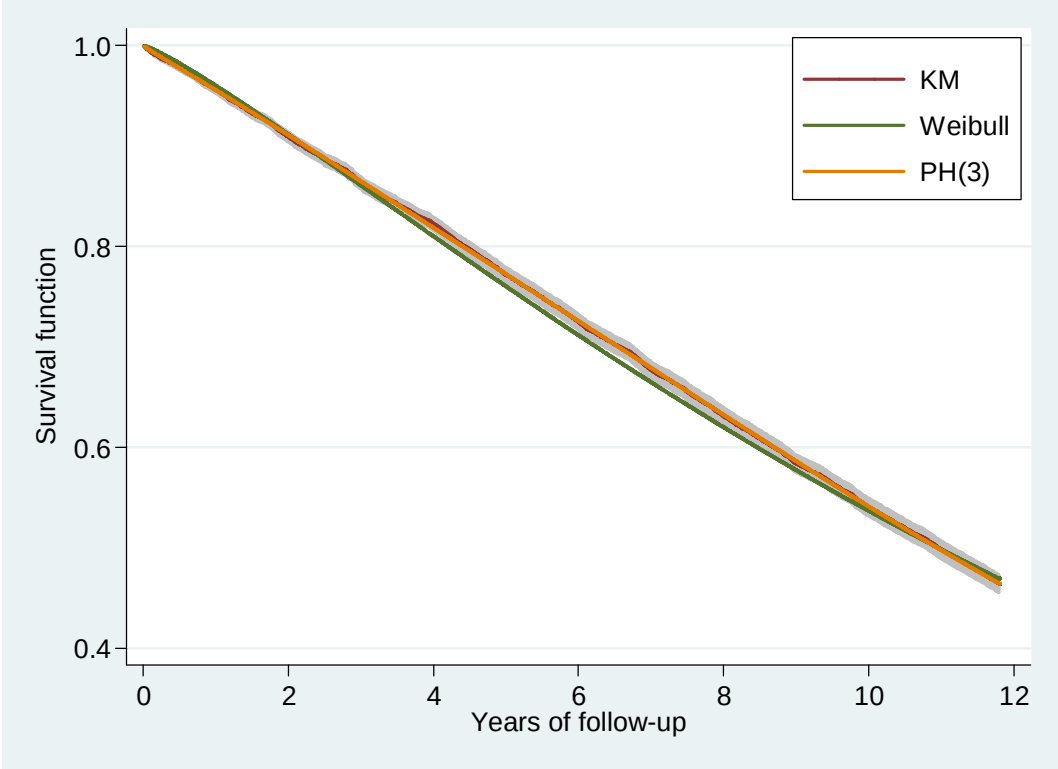

Figure 5.3. Kaplan-Meier estimator of the survival function for the population with 95% CI (shading) compared with estimates from the Weibull and PH(3) models.

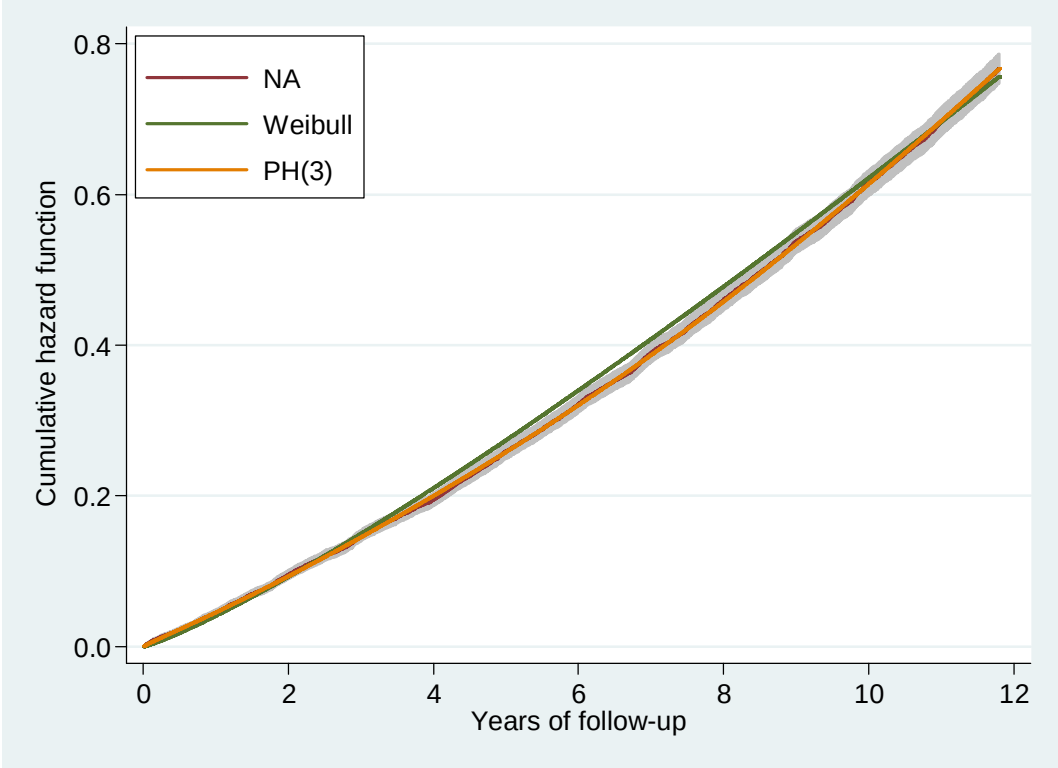

Figure 5.4. Nelson-Aelen estimator of the cumulative hazard function for the population with 95% CI (shading) compared with estimates from the Weibull and PH(3) models.

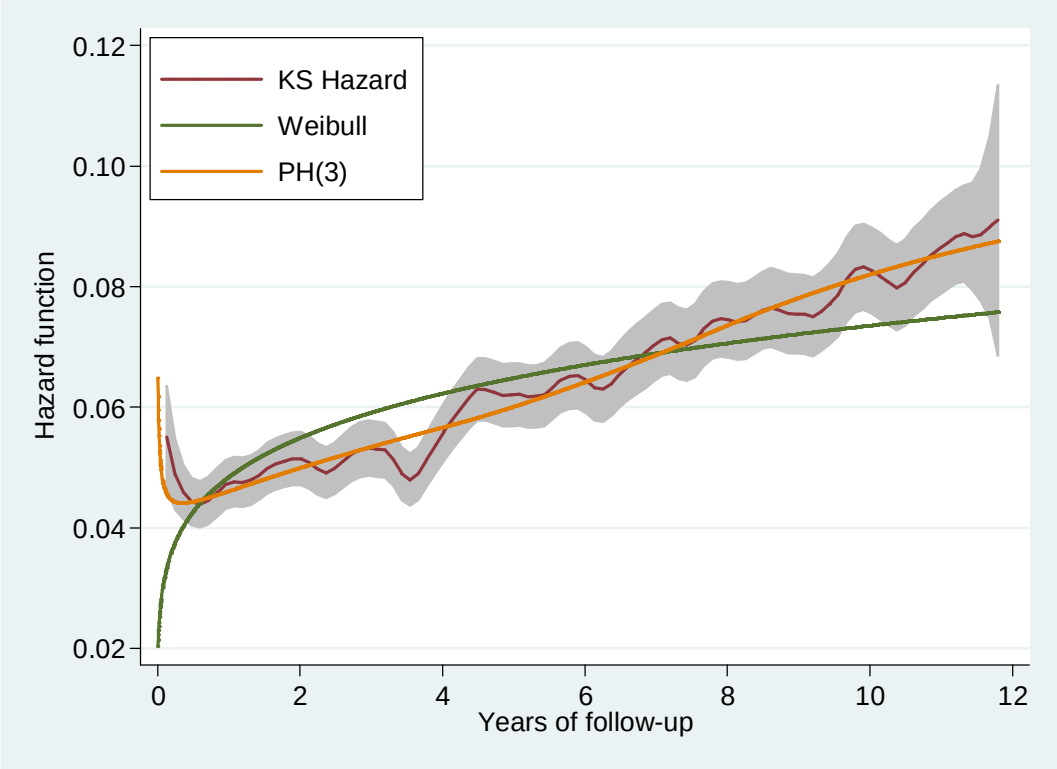

Figure 5.5. Kernel smoothed Nelson-Aalen estimator of the hazard function for the population with 95% CI (shading) compared with estimates from the Weibull and PH(3) models.

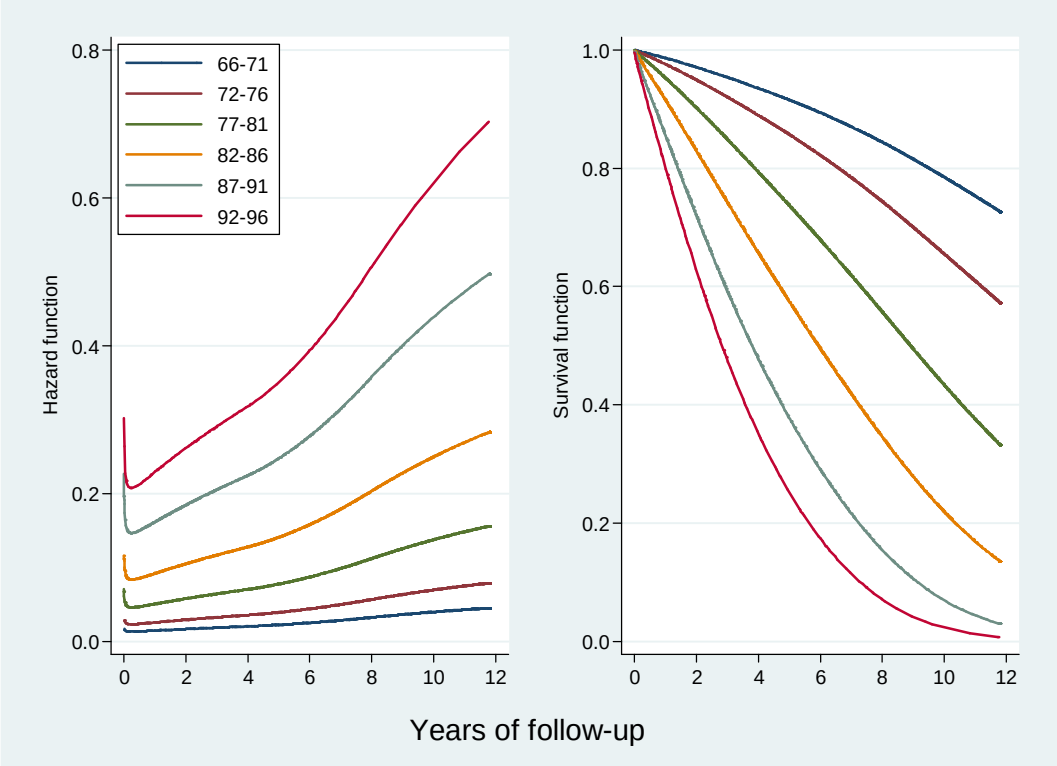

Figure 5.6. Hazard and survival functions for the different age groups estimated with PH(3) model.

To visualize the effect of age on survival and hazard the data was divided into six age groups. When we look at the hazard function in Figure 5.6 it can be seen that the hazard does not vary much for the youngest age group through the follow-up time but the change for the older groups is considerable. Substantial difference can also be seen for the

different age groups when the survival function is examined. The oldest age group has estimated survival of almost zero in the end of follow-up while the youngest age groups estimated survival is above 0,70.

In Figure 5.7 a comparison between males and females can be seen. The hazard for males is considerably higher than for females through the whole follow-up time and is in the end of follow-up around 0,10 compared with around 0,08 for females. Looking at the plot of the survival function it can be seen that males have poorer prognostics of survival than females. In the end of follow-up their estimated survival is around 0,4 compared with around 0,5 for females.

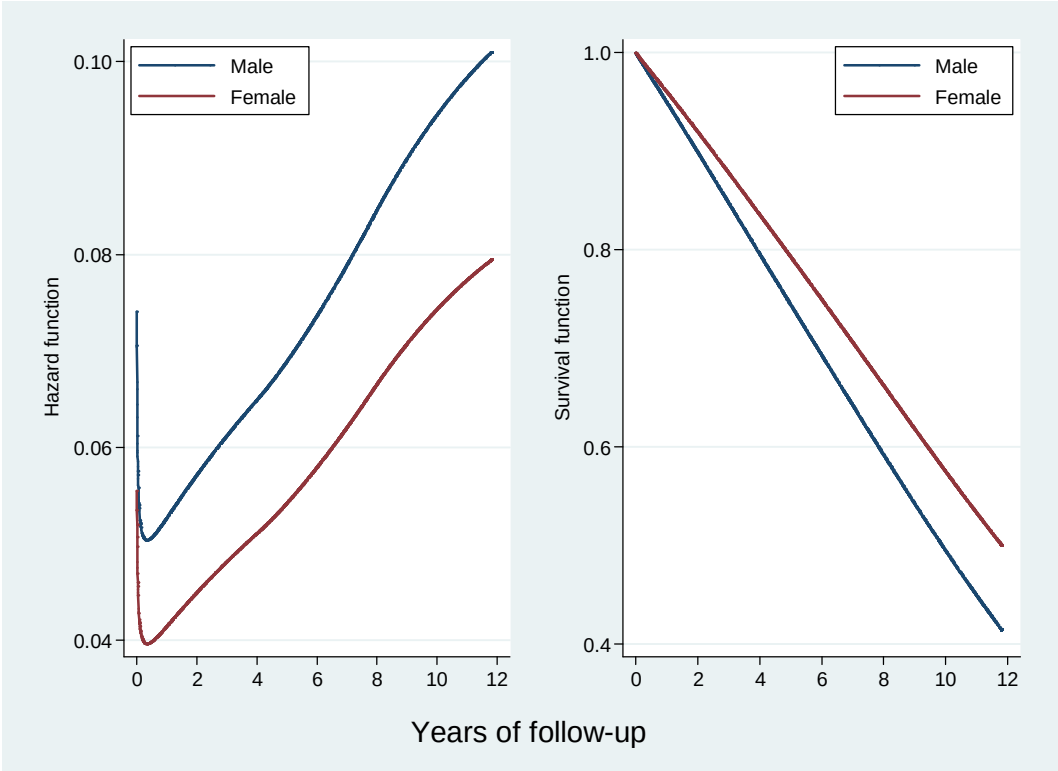

Figure 5.7. Hazard and survival functions for the males and females estimated with PH(3) model.

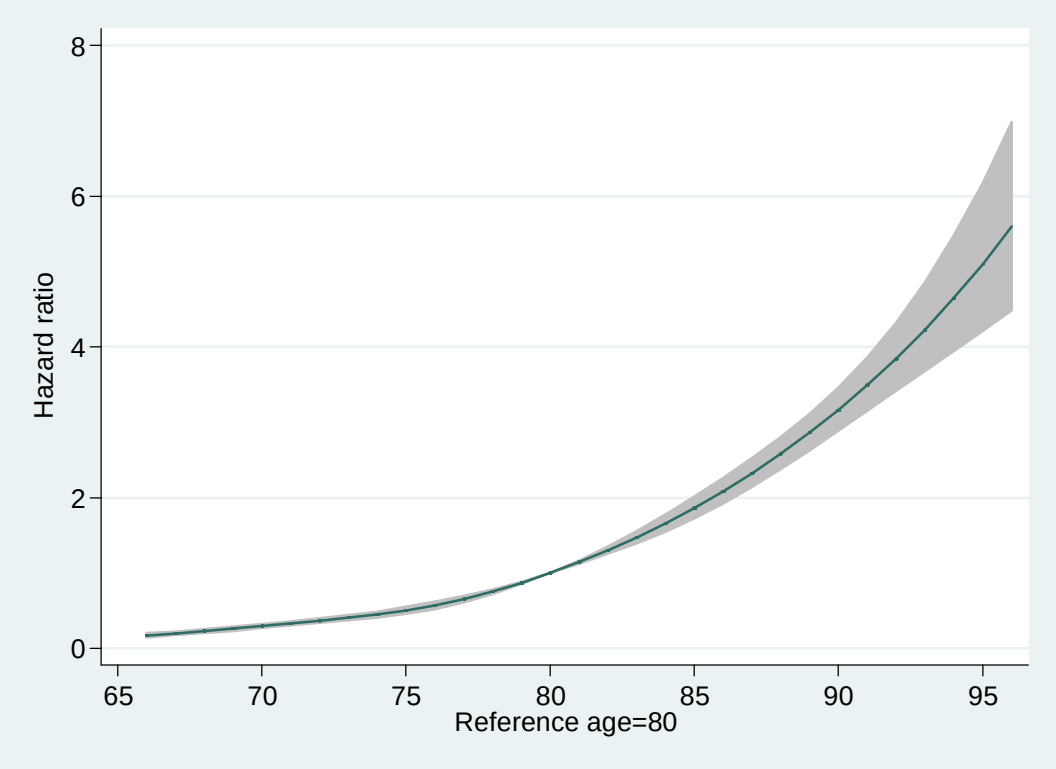

Figure 5.8. Hazard Ratio as a function of age with 95% CI. Age 80 is the reference level.

Figure 5.8 shows the hazard ratio as a function of age. Age 80 is set as the reference level and has therefore the value of 1. The confidence interval gets larger with higher age and that results from fewer observations. Individuals entering the AGES study at age 96 have over five times the hazard of experiencing the event compared with individuals entering the study at age 80.

5.3 Predictions

Figures 5.9-5.11 show some predictions that can be done by the flexible parametric models.

Figure 5.9 shows the extrapolation of the Weibull and PH(3) survival curves predicted only with data up to eight years of follow-up. The red line is the actual KM estimate until to end of follow-up. Here the predictions until to end of follow-up can be compared between the KM, Weibull and PH(3) estimates. What can be seen in the Figure is that the Weibull estimate follows the observed data well until eight years but overestimates the survival after that time. The PH(3) estimate of survival, on the other hand, follows the observed data well until the end of follow-up. The PH(3) model predicts more closely what happens after the eight year follow-up although what happens with the data beyond the 11,8 years of actual follow-up is not known.

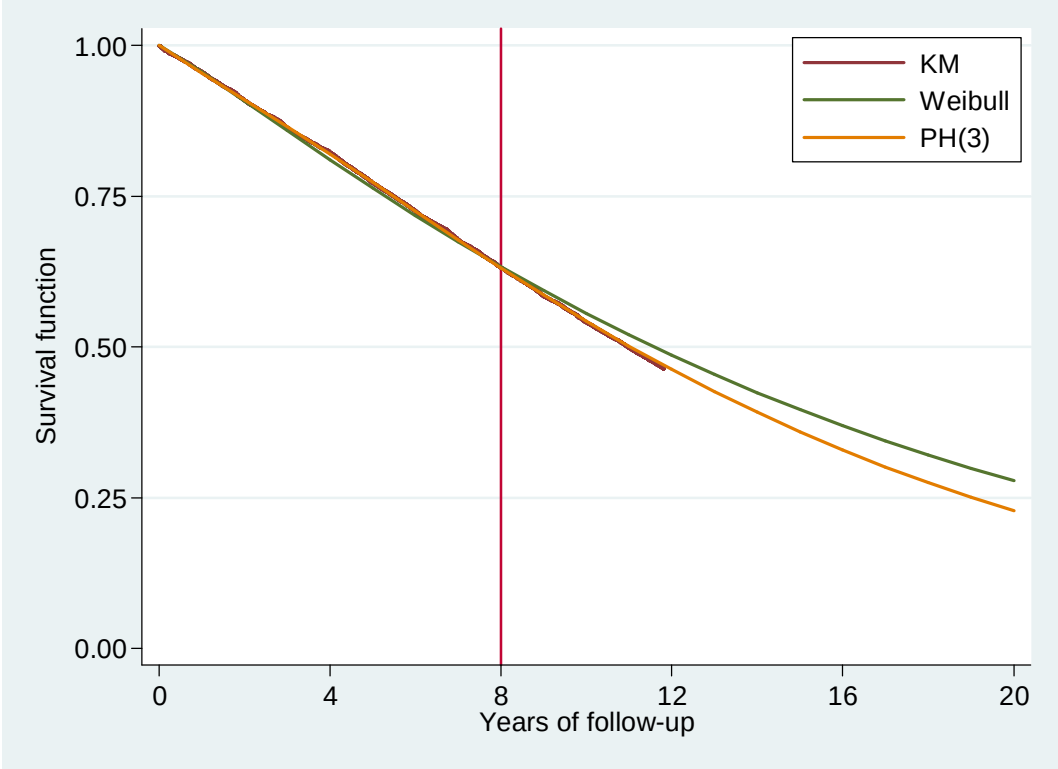

Figure 5.9. Extrapolation of survival curves up to 20 years.

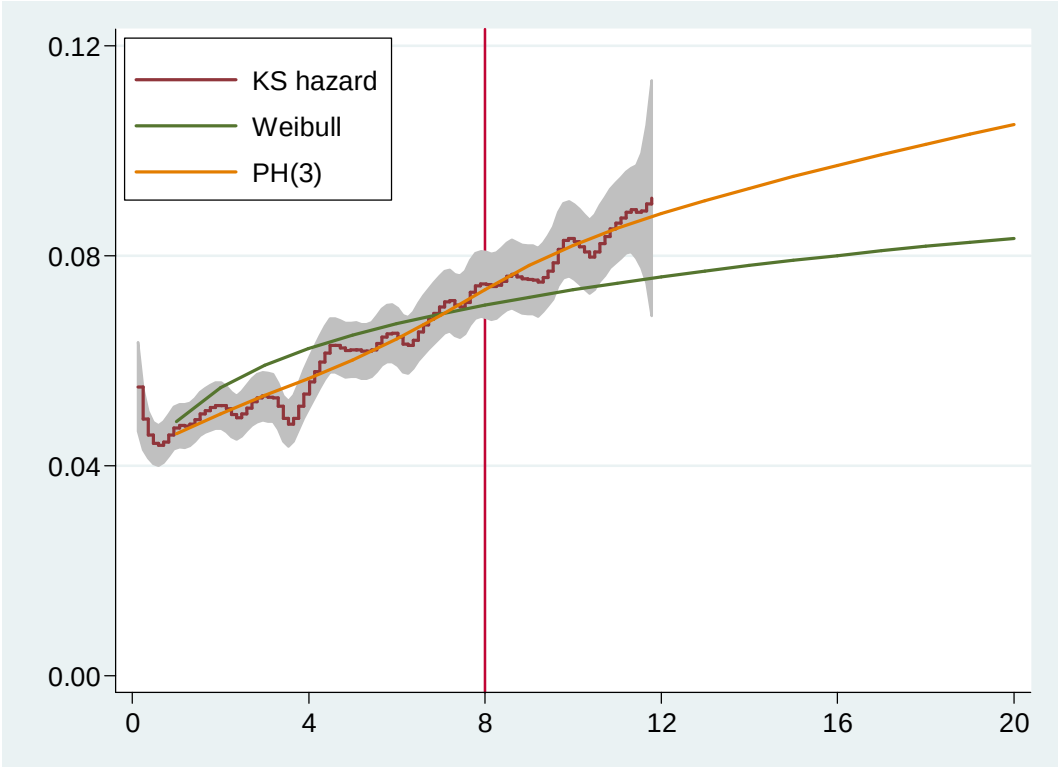

Figure 5.10. Extrapolation of the hazard function up to 20 years. The shaded area is the 95% CI for the Kernel smoothed hazard function.

Figure 5.10 shows the extrapolation of the Weibull and PH(3) hazard functions predicted with data up to eight years of follow-up. As with the predicted survival curve, the PH(3) estimate follows the observed hazard better than the Weibull estimate, both before the end of the eight year follow-up and beyond that point. However there might be some hint

that the PH(3) estimates is starting to part from the observed data by the end of the 11,8 years. It is nevertheless still within the confidence interval of the observed data.

In Figure 5.11 the mean predicted remaining survival time can be seen predicted by the PH(3) model for 10, 20, 30 and 40 years. The mean predicted survival time for the oldest people is around two years in all the figures and it does not make a difference if the predictions are done for the next 10 or 40 years. There is a difference in all the graphs between males and females as would be expected and the mean predicted survival time shortens with age, also as expected.

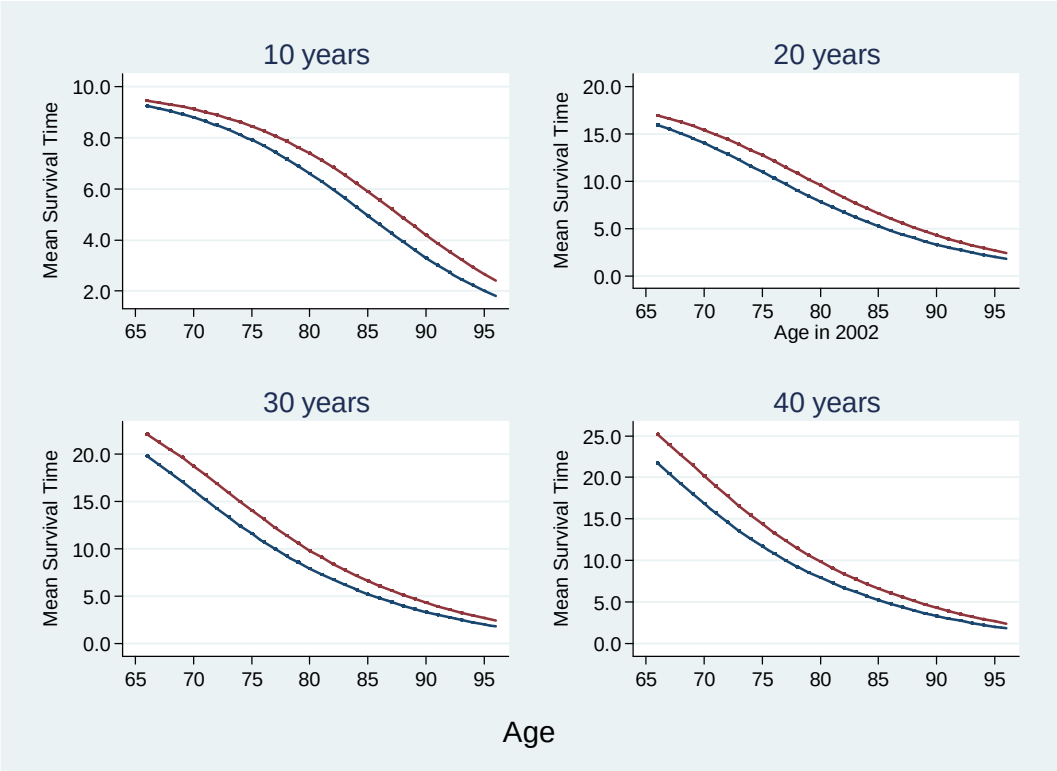

Figure 5.11. Mean predicted survival for 10, 20, 30 and 40 years by age and sex from PH(3). Solid line for males and dash/dot line for females.

5.4 Comparison to official estimations

Statistics Iceland (2014) produces national period life tables and estimates life expectancy by age and sex. In Figure 5.12 the official estimates are compared with the mean remaining survival time estimates from the PH(3) model. The population estimated by Statistics Iceland is people aged 66-94 in the years 2001-2005 but the data for the PH(3) model is people from the AGES database, aged 66-96 in the year 2002. There is considerable difference in the predictions between the estimates for the younger age groups, where the PH(3) model predicts longer remaining mean survival time for the AGES population than Statistics Iceland does for the Icelandic population. After age 80 the estimates of the AGES data modeled with PH(3) agree well with the period table from Statistics Iceland, although it seems to estimate a slightly shorter survival time.

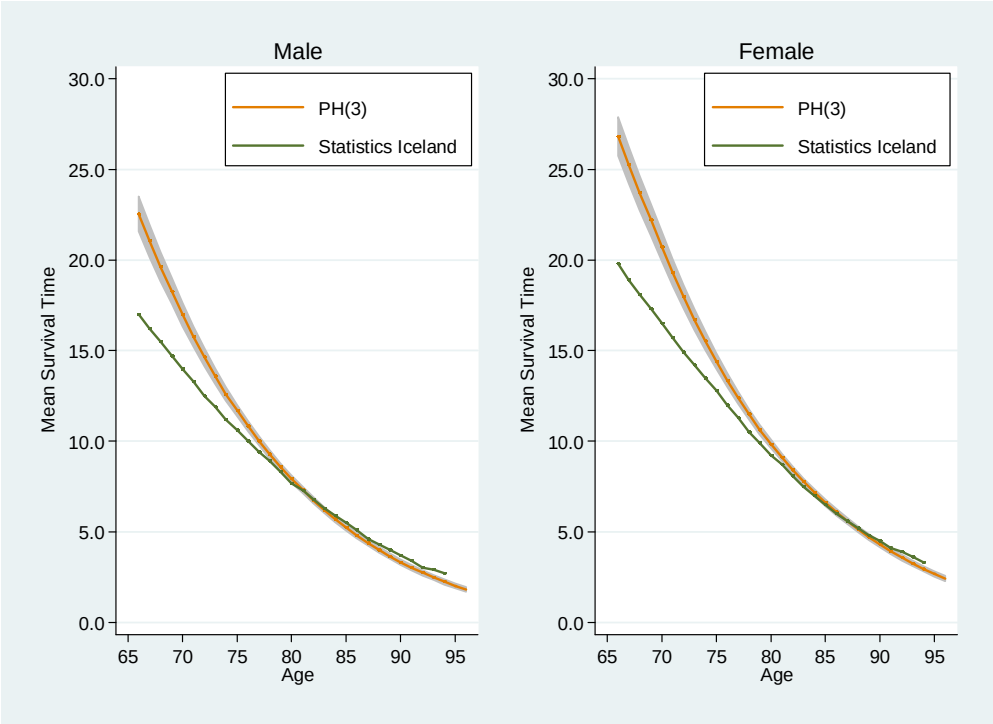

Figure 5.12. Mean predicted survival times compared between official estimates of Statistics Iceland and estimates from the PH(3) model.

The predictions from the PH(3) model can also be done with time restrictions. Instead of unrestricted as in Figure 5.12. In Figure 5.13 predictions have been restricted to 20, 25 and 30 years and plotted with the official estimates from Statistics Iceland. As can be seen the closest fit for the younger age groups seems to be between restrictions of 20 and 25 years .

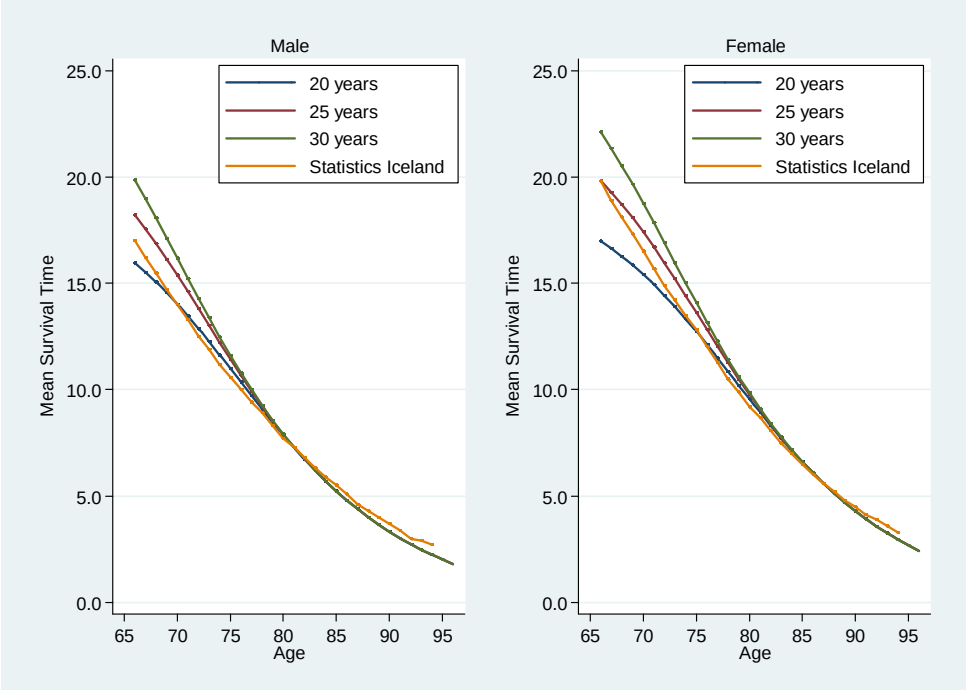

Figure 5.13. Mean predicted survival times compared between official estimates of Statistics Iceland and estimates from the PH(3) model restricted on 20, 25 and 30 years.

6 Discussion and Concluding Remarks

This paper is a modest contribution to the ongoing discussion about the possibilities available to model survival data. The main purpose of the paper was to evaluate the different methods of survival analysis and apply to the given data. The goal was to estimate if there would be a considerable improvement in fit of the data by modeling with a flexible parametric model instead of a standard parametric model. The optimal model should have hazard ratio estimates in agreement with the estimates from the Cox model and additionally the characteristics of parametric models. In that way the aim was to explore if the data could be modeled in a way that would combine the best of the two standard modeling techniques.

The results indicate that the best parametric model for the data is the flexible proportional hazard model with three degrees of freedom, PH(3). That model has two internal knots in the spline function. The AIC and BIC for the flexible model of choice indicated a much better fit than the standard Weibull model. Modeling the survival, hazard and cumulative hazard functions of both the Weibull model and the PH(3) model and comparing the plotted functions to the actual observed data indicated that the PH(3) model fits the data better than the Weibull model. Even the hazard function, that has usually been found to be less stable and harder to estimate, seems modeled appropriately by the PH(3) model while the Weibull hazard estimate has a different shape. The observed data and the different predicted functions from PH(3) therefore agree very well. The generally close agreement can be seen as crude confirmation of a good model fit. Figures 5.9 and 5.10 show that when modeling the data with information of only 8 years of follow-up and predicting beyond that point, the PH(3) model is in better agreement than the Weibull model, with what can be seen from the data until the end of real follow-up.

The comparison of the PH(3) model and the Cox model is more difficult, since there is no direct way of evaluation. The Harrel's C, representing the proportion of all usable subject pairs in which the predictions and outcomes are in agreement is the same for both models. The chi-square test of calibration also shows that both models fit the data well. Both models seem to be suitable. The main argument for choosing the flexible parametric model over the Cox model lies in what features the PH(3) model contains additionally to the Cox model, mainly in the area of predictions. The main difference of the two models is that the flexible parametric model estimates the baseline hazard while the Cox modeling technique leaves it out as nuisance factor, and therefore predictions are easier and more accurately obtained by the flexible parametric model.

The comparison of the predicted life expectancy between the official estimates of the Icelandic population from Statistics Iceland and the PH(3) predictions for the AGES population show that for the younger age groups the unrestricted PH(3) predicts much longer survival than Statistics Iceland does. It is interesting to see this difference but the dissimilarities in both populations and methods must be considered. It might be possible that the AGES population is in a better physical shape than the average Icelandic population since only people fit enough to come for examination was included in the AGES database. It must also be kept in mind that the follow-up time is only around 11,8

years and therefore it might be questionable to predict survival for many decades. Figure 5.13 shows predictions restricted to 20, 25 and 30 years and the best fit seems to be between 20 and 25 years. What happens for this group in the future is though still unknown.

The main limitation of this study is that the models have not been validated externally, that is, the model fitted to the given data has not been applied to other data to assess the fit. The fit has only been examined within the modeled data. The question of external model fit is thus unanswered.

The flexible proportional hazard model applied in this study showed a good internal fit of the given data. This is consistent with findings both of the researchers that developed the model and others that have applied it (Royston & Parmar, 2002; Lambert & Royston, 2009; Royston & Lambert, 2011; Andersson, 2013; Crowther & Lambert, 2014). The hazard function of the data in this study was not of particularly complicated shape and still there was a considerable improvement in fit with the PH(3) model compared with the Weibull. The data included information on almost 12 years of follow-up for people aged 66-96 in the beginning of the study. For people that age, not controlling for any special medical condition, the hazard should be monitonically increasing along the follow-up time, since people are getting older. In many medical data the hazard function can be much more complicated where patients are e.g. in much hazard before and around the time of surgery, then if surgery is successful the hazard decreases but might increase again if there is a risk of recurrence. More improvement of modeling data with multi-modal hazard functions should therefore be expected with the flexible parametric models compared with the standard ones.

In the case of the data modeled in this study the appropriate scale to model on is the proportional hazard. That is not true for all data and some data are better modeled under the assumption of proportional odds or the probit assumption (Royston & Lambert, 2011). In that case there is no other model equivalent to the Cox model for proportional hazard to be compared with. The advantage of being able to model flexibly with splines is then apparent, since the only real competitors are the standard parametric models, known for a frequent lack of fit.

The lack of tools for direct comparison between the Cox model and the flexible PH model is unforunately evident, at least at the moment. Hopefully research will continue on that field as flexible parametric modeling will become more common.

Hjort said in 1992 that an adequate parametric version of the Cox model would lead to more precise estimation of survival probabilities and contribute to a better understanding of the phenomenon under study. It looks like the flexible parametric hazard model might be the answer to his request. Based on the results of this paper it can be concluded that in the case of the data from the AGES study the PH(3) model fits the data at least as good as the Cox model and additionally offers the valuable predicting features of parametric modeling.

References

Allison, P., 2010. *Survival Analysis Using SAS. A Practical Guide.* Cary, USA: SAS Institute Inc.

Andersson, T. M., 2013. *Quantifying cancer patient survival; extensions and applications of cure models and life expectansy estimation.* s.l.:Ph.d Thesis. Karolinska Institute.

Cleves, M., Gutierrez, R., Gould, W. & Marchenko, Y., 2002. *An introduction to Survival Analysis Using Stata.* 3rd ed. College Station, Texas: Stata Press.

Collett, D., 2003. *Modelling Survival Data in Medical Research.* s.l.:Chapman & Hall/CRC.

Cox, D. R., 1972. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological),* 34(2), pp. 187-220.

Crowther, M. & Lambert, P., 2014. A general framework for parametric survival analysis. *Statistics in Medicine,* pp. 5280-5297.

Demler, P. V., Paynter, N. P. & Cook, N. R., 2015. Tests of calibration and goodness-of-fit in the survival setting. *Statistics in Medicine,* Issue 34, pp. 1659-1680.

Harris, T. et al., 2007. Age, Gene/Environment Susceptibility–Reykjavik Study: Multidisciplinary Applied Phenomics. *American Journal of Epidemiology,* pp. 1076-1087.

Hjort, N., 1992. On inference in parametric survival data models. *International Statistical review,* Issue 60, pp. 355-387.

Hosmer, D. & Lemeshow, S., 1980. A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics,* Issue A10, pp. 1043-1069.

Jackson, A. et al., 2011. Validation and use of a parametric model for projecting cystic fibrosis survivorship beyond observed data: a birth cohort analysis. *Thorax,* Issue 66, pp. 674-679.

Lambert, P. & Royston, P., 2009. Further development of flexible parametric models for survival analysis. *The Stata Journal,* pp. 265-290.

Office for National Statistics, 2014. *National Life Tables, United Kingdom, 2011-2013.* [Online]. Available at: http://www.ons.gov.uk/ons/rel/lifetables/national-life-tables/2011-2013/stb-uk-2011-2013.html#tab-Introduction [Accessed 06 05 2015].

Royston, P. & Lambert, P. C., 2011. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model.* College Station, Texas: A Stata Press Publication.

Royston, P. & Parmar, M., 2002. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine,* pp. 2175-2197.

Royston, P. & Sauerbrei, W., 2004. A new measure of prognostic seperation in survival data. *Statistics in Medicine,* 23(5), pp. 723-748.

Statistics Iceland, 2014. *Statistics Iceland.* [Online]. Available at: http://www.hagstofa.is/ [Accessed 04 05 2015].

Appendix A: Deriving the ln(hazard) function in flexible parametric modeling

From (Royston & Lambert, 2011).

Given the hazard function and the log hazard function

$$h(t|\mathbf{x}) = \frac{ds(\ln t\,|\gamma)}{dt}\exp\{s(\ln t|\gamma) + \mathbf{x}\boldsymbol{\beta}\}$$

$$\ln h(t|\mathbf{x}) = \ln\left\{\frac{ds(\ln t\,|\gamma)}{dt}\right\} + s(\ln t|\gamma) + \mathbf{x}\boldsymbol{\beta}$$

Evaluating the derivative

$$\frac{ds(\ln t\,|\gamma)}{dt} = \frac{d}{d\ln t}(\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \gamma_{m+1} z_{m+1})$$

$$= \gamma_1 + \gamma_2\frac{dz_2(\ln t)}{d\ln t} + \cdots + \gamma_{m+1}\frac{dz_{m+1}(\ln t)}{d\ln t}$$

Given that for $j = 2, \ldots, m+1$

$$z_j(\ln t) = \left(\ln t - k_j\right)_+^3 - \lambda_j(\ln t - k_{min})_+^3 - \left(1-\lambda_j\right)(\ln t - k_{max})_+^3$$

Then

$$\frac{dz_j}{d\ln t} = z_j' = 3\left(\ln t - k_j\right)_+^2 - 3\lambda_j(\ln t - k_{min})_+^2 - 3\left(1-\lambda_j\right)(\ln t - k_{max})_+^2$$

Finally

$$\ln h(t|\mathbf{x}) = -\ln t + \ln(\gamma_1 + \gamma_2 z_2' + \cdots + \gamma_{m+1} z_{m+1}') + \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \cdots$$
$$+\gamma_{m+1} z_{m+1} + \mathbf{x}\boldsymbol{\beta}$$

Appendix B: Comparison of functions between the Weibull and PH(3) for males and females separately

Males

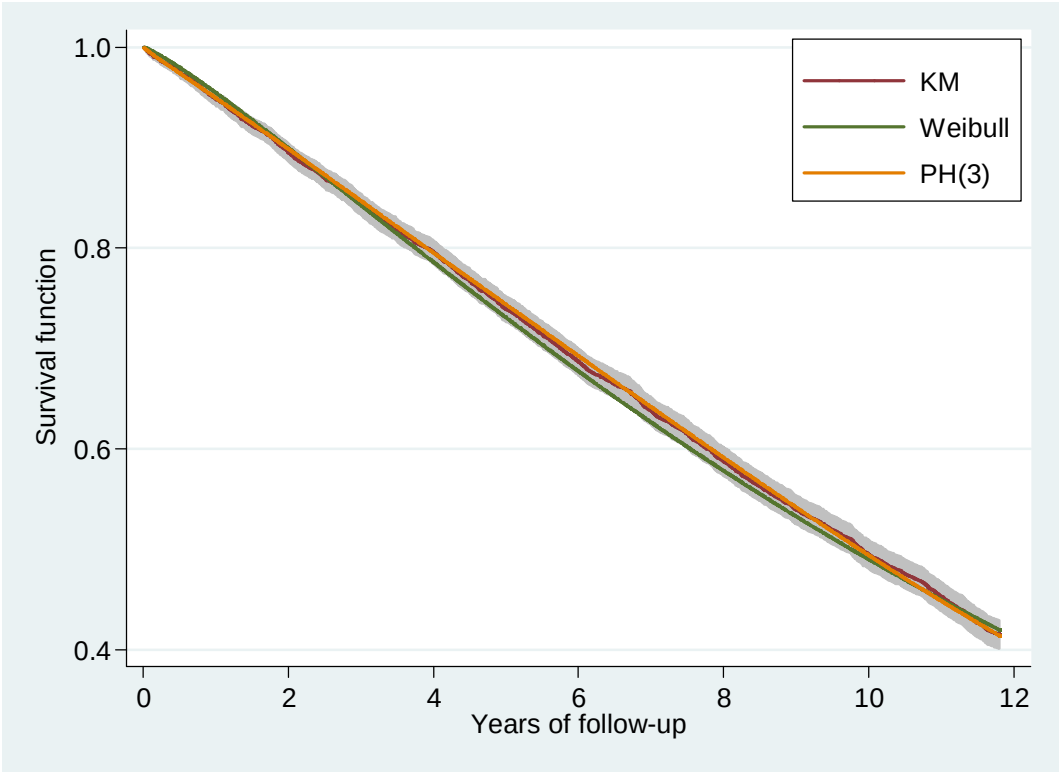

Figure B.1 Kaplan-Meier estimator of the survival function for males with 95% CI (shading) compared with estimates from the Weibull and PH(3) models.

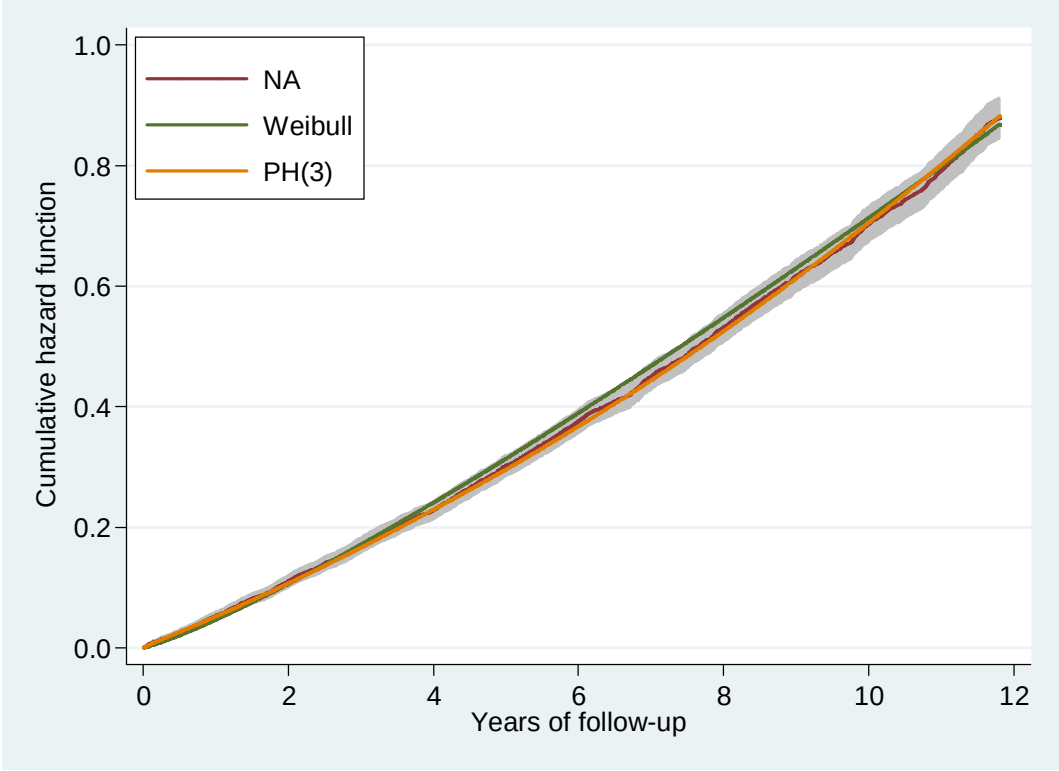

Figure B.2 Nelson-Aelen estimator of the cumulative hazard function for males with 95% CI (shading) compared with estimates from the Weibull and PH(3) models.

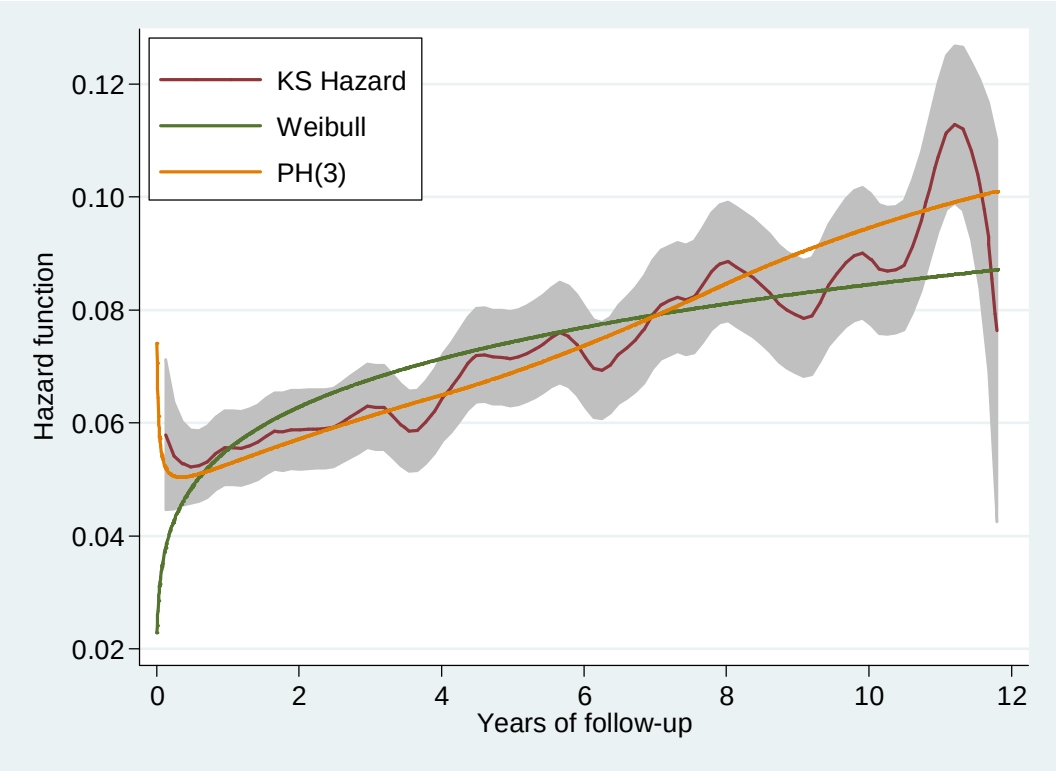

Figure B.3 Kernel smoothed Nelson-Aalen estimator of the hazard function for males with 95% CI (shading) compared with estimates from the Weibull and PH(3) models.

Females

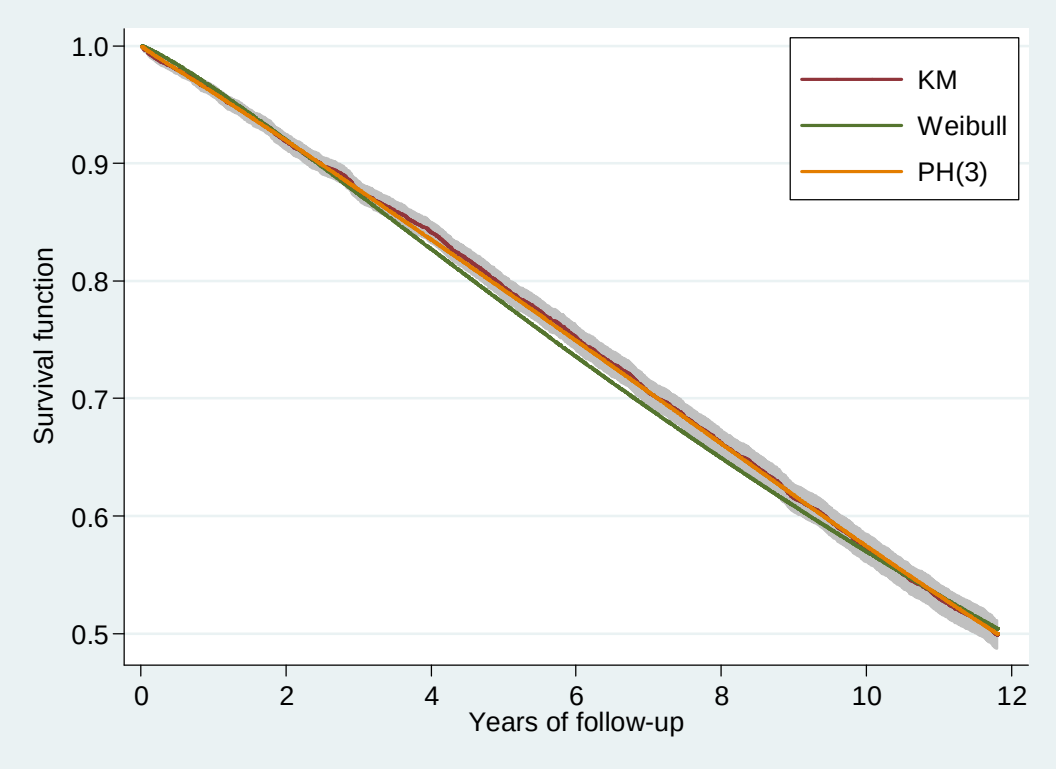

Figure B.4 Kaplan-Meier estimator of the survival function for females with 95% CI (shading) compared with estimates from the Weibull and PH(3) models.

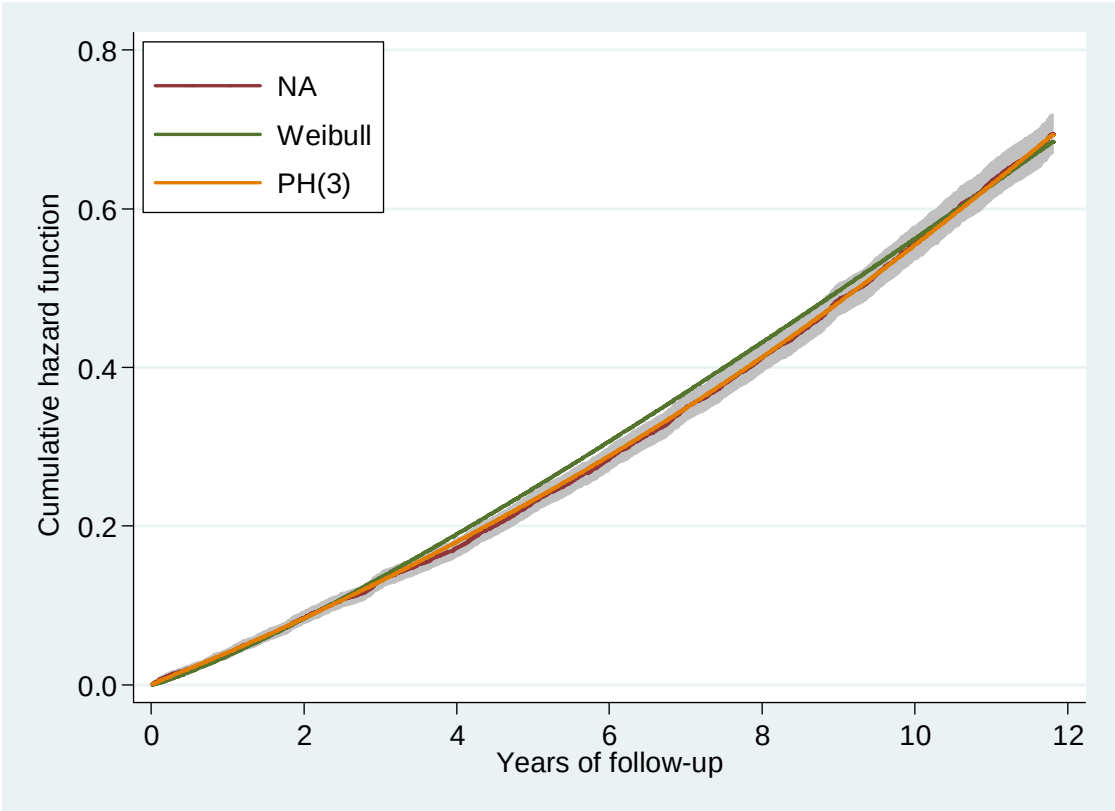

Figure B.5 Nelson-Aelen estimator of the cumulative hazard function for females with 95% CI (shading) compared with estimates from the Weibull and PH(3) models.

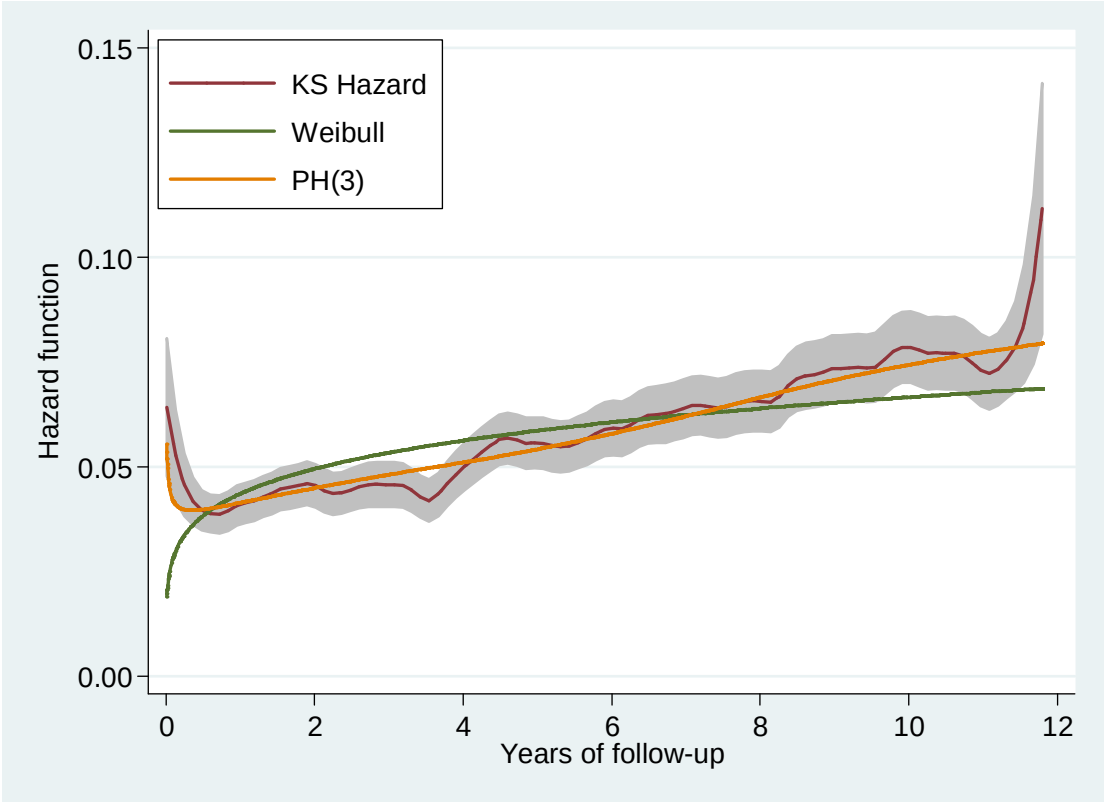

Figure B.6 Kernel smoothed Nelson-Aalen estimator of the hazard function for females with 95% CI (shading) compared with estimates from the Weibull and PH(3) models.

Appendix C: Stata codes

```
clear
// Import data //
use "C:\Users\siggigusti\Dropbox\Master\Stata\data2.dta", clear
destring, replace
// Set dependent variable and censor variable //
stset fu02y, failure(death==1)

// Information for Table 4.1 //
summarize agebegin02 lifetime fu02y, detail

// Information for Table 4.2 //
by sex, sort : inspect sex
by death, sort : inspect death

// Information for Table 5.1 //

// Recode age into groups //
recode agebegin02 (min/71=0)(72/76=1)(77/81=2)(82/86=3)(87/91=4) (92/max=5), gen(age_groups)

by sex age_groups, sort : summarize death if death==1, detail
by sex age_groups, sort : summarize death if death==0, detail

total fu02y, over(sex age_groups)

// Figure 5.1 //
sts generate sf=s if(sex==2)
sts generate sm=s if(sex==1)
line sm sf _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "Males") label(2 "Females") ring(0)
pos(1) col(1)) xlabel(0 (2) 12) xscale(r(0 12)) ytitle("Survival function") xtitle("") yla(, angle(h)
format(%5.1f)) name(g1,replace)

sts generate s70=s if(agebegin02==70)
sts generate s80=s if(agebegin02==80)
sts generate s90=s if(agebegin02==90)
line s70 s80 s90 _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "70") label(2 "80") label(3 "90")
ring(0) pos(1) col(1)) xlabel(0 (2) 12) xscale(r(0 12)) ytitle("Survival function") xtitle("") yla(, angle(h)
format(%5.1f)) name(g2,replace)

graph combine g1 g2, b2title("Years of follow-up")


sum agebegin02, meanonly
gen agec = agebegin02-r(mean)


 // Log Rank tests //
sts test sex
sts test agec
sts test agec sex

// Table 5.2 //
foreach scale in hazard odds normal{
        display _n "Scale = `scale"'
        forvalues j=1/6{
                quietly xi:stpm2 i.sex agec, df(`j') scale(`scale')
                display "df = `j', AIC = "%13.7f e(AIC) " BIC = " %13.7f e(BIC)
        }
}
```

```
// Table 5.3 //
stpm2 agec i.sex, scale(hazard) df(1) eform nolog
stpm2 agec i.sex, scale(hazard) df(2) eform nolog
stpm2 agec i.sex, scale(hazard) df(3) eform nolog
stpm2 agec i.sex, scale(hazard) df(4) eform nolog
stpm2 agec i.sex, scale(hazard) df(5) eform nolog
stpm2 agec i.sex, scale(hazard) df(6) eform nolog
stcox agec i.sex, nolog

// Check PH assumptions //
stcox agec i.sex
estat phtest, detail

// Figure 5.2 //
sts graph, hazard by(sex) noboundary yscale(log) legend(label (1 "Males") label(2 "Females") ring(0) pos(10) col(1)) xlabel(0 (2) 12) xscale(r(0 12)) ytitle("Smoothed hazard function") xtitle("Years of follow-up") yla(, angle(h) format(%7.2f)) title("") name(g2,replace)

// Table 5.4 //
foreach scale in hazard{
        display _n "Scale = `scale'"
        forvalues j=1/6{
                xi:str2d: stpm2 i.sex agec, df(`j') scale(`scale')
        }
}
// Harrell's C //
stpm2 agec i.sex, scale(hazard) df(3) nolog
stcstat2

stcox agec sex
estat concordance

// Table 5.5 //
// PH(3) //
clear
use "C:\Users\siggigusti\Dropbox\Master\Stata\data2.dta", clear
destring, replace
// Set dependent variable and censor variable //
stset fu02y, failure(death==1)
sum agebegin02, meanonly
gen agec = agebegin02-r(mean)

stpm2 i.sex agec, scale(hazard) df(3)
replace _t=11
predict xb, xbnobaseline
predict failure, failure

// Cox //
clear
use "C:\Users\siggigusti\Dropbox\Master\Stata\data2.dta", clear
destring, replace
// Set dependent variable and censor variable //
stset fu02y, failure(death==1)

sum agebegin02, meanonly
gen agec = agebegin02-r(mean)
stcox i.sex agec
predict xb, xb
su xb
```

```
predict sa, basesurv
gen st11 = sa if _t>10.99 & _t<=11
sum st11, meanonly
replace st11 = r(mean)
gen sa11 = st11^exp(xb)
gen failure = 1-sa11
label var sa11 "stcox prediction at time 11"
label var failure "risk prediction at time 11"

// Weibull //
clear
use "C:\Users\siggigusti\Dropbox\Master\Stata\data2.dta", clear
destring, replace
// Set dependent variable and censor variable //
stset fu02y, failure(death==1)
sum agebegin02, meanonly
gen agec = agebegin02-r(mean)

stpm2 i.sex agec, scale(hazard) df(1)
replace _t=11
predict xb, xbnobaseline
predict failure, failure


// Figures 5.3 - 5.5 //
// Hazard(3) //
stpm2, scale(hazard) df(3) eform
predict H, cumhazard
predict S, survival
predict h, hazard
// Weibull //
stpm2, scale(hazard) df(1) eform
predict HW, cumhazard
predict SW, survival
predict hW, hazard

//KM Survival//
sts generate ss=s
sts generate sse=se(s)
generate lo = ss - 1.96*sse
generate hi = ss + 1.96*sse
//NA Cumhaz//
sts generate hh=na
sts generate hhse=se(na)
generate lohh = hh - 1.96*hhse
generate hihh = hh + 1.96*hhse
//KS Hazard//
sts graph, hazard ci kernel(epan2) outfile(haz, replace)
append using "C:\Users\siggigusti\Dropbox\Master\stata\haz.dta"
generate loh=hazard-1.96*sqrt(Vhazard)
generate hih=hazard+1.96*sqrt(Vhazard)

//Plots//
twoway(rarea lo hi _t, pstyle(ci) sort) line ss SW S _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "") label(2 "KM") label(3 "Weibull") label(4 "PH(3)") ring(0) pos(1) col(1)) ytitle("Survival function") xtitle("Years of follow-up") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.1f)) name(g2,replace)
twoway(rarea lohh hihh _t, pstyle(ci) sort) line hh HW H _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "95% CI") label(2 "NA") label(3 "Weibull") label(4 "PH(3)") ring(0) pos(10) col(1)) ytitle("Cumulative hazard function") xtitle("Years of follow-up") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.1f)) name(g4,replace)
```

```
twoway(rarea loh hih _t, pstyle(ci) sort) line hazard hW h _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "95% CI") label(2 "KS Hazard") label(3 "Weibull") label(4 "PH(3)") ring(0) pos(10) col(1)) ytitle("Hazard function") xtitle("Years of follow-up") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.2f)) name(g3,replace)

// Figure 5.6 //
stpm2 i.age_groups, scale(hazard) df(3) eform
predict S0, survival if(age_group==0)
predict S1, survival if(age_group==1)
predict S2, survival if(age_group==2)
predict S3, survival if(age_group==3)
predict S4, survival if(age_group==4)
predict S5, survival if(age_group==5)
predict h0, hazard if(age_group==0)
predict h1, hazard if(age_group==1)
predict h2, hazard if(age_group==2)
predict h3, hazard if(age_group==3)
predict h4, hazard if(age_group==4)
predict h5, hazard if(age_group==5)
//Survival functions of different age groups //
line S0 S1 S2 S3 S4 S5 _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(off) ytitle("Survival function") xtitle("") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.1f)) name(gg1,replace)
//Hazard functins of males and females//
line h0 h1 h2 h3 h4 h5 _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "66-71") label(2 "72-76") label(3 "77-81") label(4 "82-86") label(5 "87-91") label(6 "92-96") ring(0) pos(10) col(1)) ytitle("Hazard function") xtitle("") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.1f)) name(gg2,replace)
//Combine//
graph combine gg2 gg1, b2title("Years of follow-up")

// Figure 5.7 //
stpm2 i.sex , scale(hazard) df(3) eform
predict Sm, survival if(sex==1)
predict Sf, survival if(sex==2)
predict hm, hazard if(sex==1)
predict hf, hazard if(sex==2)
//Survival functions of males and females//
line Sm Sf _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "Male") label(2 "Female") ring(0) pos(1) col(1)) ytitle("Survival function") xtitle("") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.1f)) name(gg1,replace)
//Hazard functions of males and females//
line hm hf _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "Male") label(2 "Female") ring(0) pos(10) col(1)) ytitle("Hazard function") xtitle("") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.2f)) name(gg2,replace)
//Combine//
graph combine gg2 gg1, b2title("Years of follow-up")

// Figure 5.8 //
rcsgen agebegin02, df(6) gen(agercs) orthog
global ageknots `r(knots)'
matrix R=r(R)
stpm2 agercs1-agercs6, scale(hazard) df(3) nolog eform

generate refage=80 in 1
rcsgen refage in 1, knots($ageknots) gen(ragercs) rmatrix(R)
local c1=ragercs1[1]
local c2=ragercs2[1]
local c3=ragercs3[1]
local c4=ragercs4[1]
local c5=ragercs5[1]
local c6=ragercs6[1]
```

```
predict hr, hrnumerator(agercs1 . agercs2 . agercs3 . agercs4 .  agercs5 . agercs6 .) hrdenominator(agercs1 `c1' agercs2 `c2' agercs3 `c3' agercs4 `c4' agercs5 `c5' agercs6 `c6') ci
twoway (rarea hr_lci hr_uci agebegin02, pstyle(ci) sort) (line hr agebegin02, sort lstyle(refline ..) pstyle(p2 ..)), leg(off) ytitle("Hazard ratio") xtitle("Reference age=80") xlabel(65 (5) 95) xscale(r(65 96)) ylab(, angle(h) format(%4.0f)) name(g2, replace)

//Figure 5.9//
clear
use "C:\Users\siggigusti\Dropbox\Master\Stata\data2.dta", clear
destring, replace
stset fu02y, failure(death==1) exit(time 8)
range timev 0 20 21

//Predict up to 20 years //
stpm2, scale(hazard) df(3)
predict s_hazard_3, survival timevar(timev)

stpm2, scale(hazard) df(1)
predict s_hazard_1, survival timevar(timev)

// Get Kaplan-Meier up to 20 years
stset fu02y, failure(death==1)
sts gen s_kap = s
sts generate sse=se(s)
generate lo = ss - 1.96*sse
generate hi = ss + 1.96*sse

// Plot //
twoway (line s_kap _t, c(J) sort lwidth(medthick ..))(line s_hazard_1 s_hazard_3 timev, sort lwidth(medthick ..)), xline(8) ylabel(0(0.25)1, angle(h) format(%4.2f)) xlabel(0(4)20) xtitle("Years of follow-up") ytitle("Survival function") legend(label(1 "KM") label(2 "Weibull") label(3 "PH(3)") ring(0) pos(1) col(1))  name(g1, replace)

//Figure 5.10 - hazard//
//Predict up to 20 years //
stpm2, scale(hazard) df(3)
predict h_hazard_3, hazard timevar(timev)

stpm2, scale(hazard) df(1)
predict h_hazard_1, hazard timevar(timev)

// Get Nelson-Aelen up to 20 years
stset fu02y, failure(death==1)
sts graph, hazard ci kernel(epan2) outfile(haz, replace)
append using "C:\Users\siggigusti\Dropbox\Master\haz.dta"
generate loh=hazard-1.96*sqrt(Vhazard)
generate hih=hazard+1.96*sqrt(Vhazard)

//plot
twoway (rarea loh hih _t, pstyle(ci) sort)(line hazard _t, c(J) sort)(line h_hazard_1 h_hazard_3 timev, sort), xline(8) ylabel(0(0.04)0.12, angle(h) format(%4.2f)) xlabel(0(4)20) xtitle("") ytitle("") legend(label(1 "") label(2 "KS hazard") label(3 "Weibull") label(4 "PH(3)") ring(0) pos(10) col(1))  name(g1, replace)

// Figure 5.11 //
stpm2 agec i.sex, scale(hazard) df(3)

predict rmstm10, rmst tmax(10) at(sex 1)
predict rmstf10, rmst tmax(10) at(sex 2)
```

```
line rmstm10 rmstf10 agebegin02, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(off) ytitle("Mean Survival Time") xtitle("Age in 2002") xlabel(65 (5) 95) xscale(r(66 96)) yla(, angle(h) format(%5.1f)) name(g1,replace)

predict rmstm20, rmst tmax(20) at(sex 1)
predict rmstf20, rmst tmax(20) at(sex 2)
line rmstm20 rmstf20 agebegin02, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(off) ytitle("Mean Survival Time") xtitle("Age in 2002") xlabel(65 (5) 95) xscale(r(66 96)) yla(, angle(h) format(%5.1f)) name(g2,replace)

predict rmstm30, rmst tmax(30) at(sex 1)
predict rmstf30, rmst tmax(30) at(sex 2)
line rmstm30 rmstf30 agebegin02, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(off) ytitle("Mean Survival Time") xtitle("Age in 2002") xlabel(65 (5) 95) xscale(r(66 96)) yla(, angle(h) format(%5.1f)) name(g3,replace)

predict rmstm40, rmst tmax(40) at(sex 1)
predict rmstf40, rmst tmax(40) at(sex 2)
line rmstm40 rmstf40 agebegin02, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(off) ytitle("Mean Survival Time") xtitle("Age in 2002") xlabel(65 (5) 95) xscale(r(66 96)) yla(, angle(h) format(%5.1f)) name(g4,replace)

graph combine g1 g2 g3 g4, b2title("Age")

// Figure 5.12 //
clear
use "C:\Users\siggigusti\Dropbox\Master\Stata\data2.dta", clear
destring, replace
// Set dependent variable and censor variable //
stset fu02y, failure(death==1)

stpm2 agec i.sex, scale(hazard) df(3)
predict rmstm, rmst tmax(50) at(sex 1) ci
predict rmstf, rmst tmax(50) at(sex 2) ci

generate hagm = .
replace hagm = 17 if agebegin02==66
replace hagm = 16.2 if agebegin02==67
replace hagm = 15.5 if agebegin02==68
replace hagm = 14.7 if agebegin02==69
replace hagm = 14 if agebegin02==70
replace hagm = 13.3 if agebegin02==71
replace hagm = 12.5 if agebegin02==72
replace hagm = 11.9 if agebegin02==73
replace hagm = 11.2 if agebegin02==74
replace hagm = 10.6 if agebegin02==75
replace hagm = 10 if agebegin02==76
replace hagm = 9.4 if agebegin02==77
replace hagm = 8.9 if agebegin02==78
replace hagm = 8.3 if agebegin02==79
replace hagm = 7.7 if agebegin02==80
replace hagm = 7.3 if agebegin02==81
replace hagm = 6.8 if agebegin02==82
replace hagm = 6.3 if agebegin02==83
replace hagm = 5.9 if agebegin02==84
replace hagm = 5.5 if agebegin02==85
replace hagm = 5.1 if agebegin02==86
replace hagm = 4.6 if agebegin02==87
replace hagm = 4.3 if agebegin02==88
replace hagm = 4 if agebegin02==89
```

```
replace hagm = 3.7 if agebegin02==90
replace hagm = 3.4 if agebegin02==91
replace hagm = 3 if agebegin02==92
replace hagm = 2.9 if agebegin02==93
replace hagm = 2.7 if agebegin02==94

generate hagf = .
replace hagf = 19.8 if agebegin02==66
replace hagf = 18.9 if agebegin02==67
replace hagf = 18.1 if agebegin02==68
replace hagf = 17.3 if agebegin02==69
replace hagf = 16.5 if agebegin02==70
replace hagf = 15.7 if agebegin02==71
replace hagf = 14.9 if agebegin02==72
replace hagf = 14.2 if agebegin02==73
replace hagf = 13.5 if agebegin02==74
replace hagf = 12.8 if agebegin02==75
replace hagf = 12 if agebegin02==76
replace hagf = 11.3 if agebegin02==77
replace hagf = 10.5 if agebegin02==78
replace hagf = 9.9 if agebegin02==79
replace hagf = 9.2 if agebegin02==80
replace hagf = 8.7 if agebegin02==81
replace hagf = 8.1 if agebegin02==82
replace hagf = 7.5 if agebegin02==83
replace hagf = 7 if agebegin02==84
replace hagf = 6.5 if agebegin02==85
replace hagf = 6 if agebegin02==86
replace hagf = 5.6 if agebegin02==87
replace hagf = 5.2 if agebegin02==88
replace hagf = 4.8 if agebegin02==89
replace hagf = 4.5 if agebegin02==90
replace hagf = 4.1 if agebegin02==91
replace hagf = 3.9 if agebegin02==92
replace hagf = 3.6 if agebegin02==93
replace hagf = 3.3 if agebegin02==94

// Plot //
twoway(rarea rmstm_lci rmstm_uci agebegin02, pstyle(ci) sort) line rmstm hagm agebegin02, sort
lpattern(l ´_ _.) lwidth(medthick ..) legend(label(1 "") label(2 "PH(3)") label(3 "Statistics Iceland") ring(0)
pos(1) col(1)) ytitle("Mean Survival Time") xtitle("Age") t2title("Male") ylabel(0 (5) 30) xlabel(65 (5) 95)
xscale(r(66 96)) yla(, angle(h) format(%5.1f)) name(g1,replace)
twoway(rarea rmstf_lci rmstf_uci agebegin02, pstyle(ci) sort) line rmstf hagf agebegin02, sort lpattern(l ´_
_.) lwidth(medthick ..) legend(label(1 "") label (2 "PH(3)") label(3 "Statistics Iceland") ring(0) pos(1)
col(1)) ytitle("Mean Survival Time") xtitle("Age") t2title("Female") ylabel(0 (5) 30) xlabel(65 (5) 95)
xscale(r(66 96)) yla(, angle(h) format(%5.1f)) name(g2,replace)
graph combine g1 g2, b2title("")

// Figure 5.13 //
stpm2 agec i.sex, scale(hazard) df(3)
predict rmstm20, rmst tmax(20) at(sex 1) ci
predict rmstf20, rmst tmax(20) at(sex 2) ci

predict rmstm25, rmst tmax(25) at(sex 1) ci
predict rmstf25, rmst tmax(25) at(sex 2) ci

predict rmstm30, rmst tmax(30) at(sex 1) ci
predict rmstf30, rmst tmax(30) at(sex 2) ci

// Plot //
```

```
line rmstm20 rmstm25 rmstm30 hagm agebegin02, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label(1 "20 years") label(2 "25 years") label(3 "30 years") label(4 "Statistics Iceland") ring(0) pos(1) col(1)) ytitle("Mean Survival Time") xtitle("Age") t2title("Male") ylabel(0 (5) 25) xlabel(65 (5) 95) xscale(r(66 96)) yla(, angle(h) format(%5.1f)) name(g3,replace)
line rmstf20 rmstf25 rmstf30 hagf agebegin02, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label(1 "20 years") label(2 "25 years") label(3 "30 years") label(4 "Statistics Iceland") ring(0) pos(1) col(1)) ytitle("Mean Survival Time") xtitle("Age") t2title("Female") ylabel(0 (5) 25) xlabel(65 (5) 95) xscale(r(66 96)) yla(, angle(h) format(%5.1f)) name(g4,replace)
graph combine g3 g4, b2title("")


// Figures B.1-B.3 //
// Hazard(3) //
stpm2 i.sex, scale(hazard) df(3) eform
predict H, cumhazard if(sex==1)
predict S, survival if(sex==1)
predict h, hazard if(sex==1)
// Weibull //
stpm2 i.sex, scale(hazard) df(1) eform
predict HW, cumhazard if(sex==1)
predict SW, survival if(sex==1)
predict hW, hazard if(sex==1)

//KM Survival//
sts generate ss=s  if(sex==1)
sts generate sse=se(s) if(sex==1)
generate lo = ss - 1.96*sse
generate hi = ss + 1.96*sse
//NA Cumhazard//
sts generate hh=na if(sex==1)
sts generate hhse=se(na) if(sex==1)
generate lohh = hh - 1.96*hhse
generate hihh = hh + 1.96*hhse
//KS Hazard//
drop if(sex==2)
sts graph, hazard ci kernel(epan2) outfile(haz, replace)
append using "C:\Users\siggigusti\Dropbox\Master\haz.dta"
generate loh=hazard-1.96*sqrt(Vhazard)
generate hih=hazard+1.96*sqrt(Vhazard)

//Plots//
twoway(rarea lo hi _t, pstyle(ci) sort) line ss SW S _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "95% CI") label(2 "KM") label(3 "Weibull") label(4 "PH(3)") ring(0) pos(1) col(1)) ytitle("Survival function") xtitle("Years of follow-up") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.1f)) name(g2,replace)
twoway(rarea loh hih _t, pstyle(ci) sort) line hazard hW h _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "95% CI") label(2 "KS Hazard") label(3 "Weibull") label(4 "PH(3)") ring(0) pos(10) col(1)) ytitle("Hazard function") xtitle("Years of follow-up") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.2f)) name(g3,replace)
twoway(rarea lohh hihh _t, pstyle(ci) sort) line hh HW H _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "95% CI") label(2 "NA") label(3 "Weibull") label(4 "PH(3)") ring(0) pos(10) col(1)) ytitle("Cumulative hazard function") xtitle("Years of follow-up") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.1f)) name(g4,replace)

// Figures B.4-B.5 //
// Hazard(3) //
stpm2 i.sex, scale(hazard) df(3) eform
predict H, cumhazard if(sex==2)
predict S, survival if(sex==2)
predict h, hazard if(sex==2)
```

```
// Weibull //
stpm2 i.sex, scale(hazard) df(1) eform
predict HW, cumhazard if(sex==2)
predict SW, survival if(sex==2)
predict hW, hazard if(sex==2)

//KM Survival//
sts generate ss=s  if(sex==2)
sts generate sse=se(s) if(sex==2)
generate lo = ss - 1.96*sse
generate hi = ss + 1.96*sse
//NA Cumhazard//
sts generate hh=na if(sex==2)
sts generate hhse=se(na) if(sex==2)
generate lohh = hh - 1.96*hhse
generate hihh = hh + 1.96*hhse
//KS Hazard//
drop if(sex==1)
sts graph, hazard ci kernel(epan2) outfile(haz, replace)
append using "C:\Users\siggigusti\Dropbox\Master\haz.dta"
generate loh=hazard-1.96*sqrt(Vhazard)
generate hih=hazard+1.96*sqrt(Vhazard)

//Plots//
twoway(rarea lo hi _t, pstyle(ci) sort) line ss SW S _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "95% CI") label(2 "KM") label(3 "Weibull") label(4 "PH(3)") ring(0) pos(1) col(1)) ytitle("Survival function") xtitle("Years of follow-up") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.1f)) name(g2,replace)
twoway(rarea loh hih _t, pstyle(ci) sort) line hazard hW h _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "95% CI") label(2 "KS Hazard") label(3 "Weibull") label(4 "PH(3)") ring(0) pos(10) col(1)) ytitle("Hazard function") xtitle("Years of follow-up") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.2f)) name(g3,replace)
twoway(rarea lohh hihh _t, pstyle(ci) sort) line hh HW H _t, sort lpattern(l ´_ _.) lwidth(medthick ..) legend(label (1 "95% CI") label(2 "NA") label(3 "Weibull") label(4 "PH(3)") ring(0) pos(10) col(1)) ytitle("Cumulative hazard function") xtitle("Years of follow-up") xlabel(0 (2) 12) xscale(r(0 12)) yla(, angle(h) format(%5.1f)) name(g4,replace)
```