Semiparametric Analysis of Survival Data with Applications in Agricultural Science

by

Asheber Sewalem

A Thesis

presented to

The University of Guelph

In partial fulfilment of requirements

for the degree of

Master of Science

in

Mathematics and Statistics

Guelph, Ontario, Canada

ABSTRACT

SEMIPARAMETRIC ANALYSIS OF SURVIVAL DATA WITH APPLICATIONS IN AGRICULTURAL SCIENCE

Asheber Sewalem **Advisors:**

University of Guelph, 2012 **Professors T. Desmond & R. Singh**

This thesis explores the association between a response variable and various regressors in dairy cattle breeding data using the various survival models in general and the partially linear single index survival model (PLSISM) in particular. In this study calf survival data and culling data were used. The calf survival data contains the following information: survival time, birth weight, weaning weight, calving ease score, average daily gain, number of disease incidences and serum total protein content. The culling data include, survival time, herd size variation, production level (milk, fat and protein), type of supervision, body condition score and age at first calving. Both data sets contain herd, year and season of calving and were analyzed using the various survival models. The Weibull model, however, was used for detailed analyses of the data sets. The nonparametric vector of PLSISM includes body weight, total serum

protein and average daily gain for calf survival data and age at first calving, fat production and body condition core for culling data. The parametric vector of PLSISM consists of the rest of the covariates. The results show that the estimates of the parametric component are similar in the two models (Weibull and PLSISM). However, the estimates of the nonparametric component differ from parametric analysis. This difference may be attributed largely to the nonlinearity of the estimated function indicating the standard linear survival model does not adequately describe the underlying association between the response variable and the various covariates in this study. This is the first implementation and application of this complex model, PLSISM, with large real censored data.

Dedicated

to

My family

Acknowledgements

It gives me great pleasure to express my deepest respect and sincere thanks to my advisors Professor A.F. Desmond and R.S. Singh, for their encouragement, valuable suggestions, discussion and constructive criticism of my work.

I would also like to express my gratitude to Professor Lu for his extremely helpful guidance and advice provided to me during the course of this study and providing me, the core software that handles the PLSISM.

I wish to thank Drs Filippo Miglior and Gerrit Kistemaker for their moral support at the beginning of my study and also throughout the study period. I am also grateful to all my friends and colleagues at Department as well as my co-workers for their encouragements.

I am gratefully indebted to AAFC-CDN for providing me the opportunity to pursue this program as a part time student.

Finally I wish to express my deepest appreciation to my wife, Atsede Teferra and children Isabella, Noah and Yesmrach for their unfailing support and patience.

Table of Contents

List of Tables

List of Figures

Chapter 1

1.1 Introduction

1.1.1 Survival analysis

Survival analysis is one of the statistical methods that deal with data analysis where the dependent variables of interest are time, and the occurrence of events. In other words, it is the modeling of time to event data (Lawless, 2003; Kleinbaum and Klein, 2005).

The use of survival analysis is extensively applied in many fields such as biology (Kleinbaum, 1996), medicine (Nardi and Schemper, 2003), public health (Ector et al., 1996; Clark et al., 2003; Bradburn et al., 2003), epidemiology (Shoukri et al., 1998) and in many agricultural (Ducrocq, 2002; Sewalem et al., 2005) and financial sectors (Jarrow and Turnbull, 2000).

Generally, analyses of survival data require the modeling of time-to-event data, such as the time until death, disease incidence or some other negative individual experience and usually referred to as failure (Kleinbaum and Klein, 2005). However, these procedures have much wider applicability. They can be used, for example, to study age at marriage, the duration of

marriage, the intervals between successive births to a woman and calvings to cows, and length of life time for a particular study in which case the failure is considered as a positive event (Kleinbaum and Klein, 2005).

The time to the event of interest is called survival time which is the length of time from time of origin to the time the event of interest occurs. The survival function is a basic quantity employed to describe the probability that an individual survives beyond a specified time (Lawless, 2003).

In survival analysis, most often, the data are collected in the study population over a predetermined period of time and hence the time to event may not be observed for all individuals, which is called censored records. Censoring of records occurs when the survival time of an individual or material is incomplete due to some random causes. In other words, the data set can be exact or censored, and it may also be truncated. Exact data, also known as uncensored data, occurs when the clear-cut time until the event of interest is identified. Censored data occurs when a subject's time until the event of interest is identified only to occur in a certain period of time.

Usually there are three types of censoring. The first one is right censoring when the event occurs after the observed survival time. In this case the end point is not observed. This type of censoring may occur due to no event

before the study ends, loss of follow up during the study period or the individual may withdraw from the study for various reasons. In all cases the survival time for right censoring is less than the actual survival time of the individual in question. Let C denote the censoring time, that is, the time beyond which the study subject cannot be observed. The observed survival time is also referred to as follow up time. It starts at time 0 and continues until the event X or a censoring time C, whichever comes first. The observed data are denoted by (T, Δ), where T = min (X, C) is the follow-up time, and Δ = $I_{(X \le C)}$ is an indicator for status at the end of follow up,

$$\Delta = I_{(X \le C)} = \begin{cases} 0 \text{ if } X > C \text{ (censoring rccord)} \\ 1 \text{ if } X \le C \text{ (uncensored record)} \end{cases}$$

The other types of censoring which are less common are left censoring and interval censoring. Left censoring occurs if the researcher observes the presence of a condition but does not know where it began. In this case survival time is considered to be left censored if it is less than a censoring time C. The data observed on the individual can be recorded as (T, Δ) where

$$T = \max(X, C),$$

$$\Delta = \begin{cases} 0 \text{ if } X = C \text{ (censoring rccord)} \\ 1 \text{ if } T = X \text{ (uncensored record)} \end{cases}$$

Examples of this type of censoring include studying age at puberty in heifers. Time-to-event is the age at which a heifer reaches puberty, usually around 11-12 months of age. So left censoring occurs if heifers are inseminated before that specified period of time.

In interval censoring, the individual or the material is known to have experienced an event within an interval of time but the actual survival time is not known. The actual occurrence time of the event is known within an interval of time. Interval censoring mostly occurs in clinical trials where patients have periodic follow ups, and in industrial experiments where equipment items are inspected periodically (Lawless, 2003).

Right censoring is by far the most frequent type of censoring and hereafter, in this thesis, censoring means right censored records.

1.1.2. Survival distributions

Since survival data involves censored and uncensored records, unlike regular data set, presentation of summary of statistics is not possible. In order to summarize the data, therefore, the underlying distribution has to be estimated. One can then obtain the summary of the data such as the mean and standard deviation of the data.

Survival time data quantify the time to a certain event. These times are subject to random variations, and like any random variables, form a distribution. Let T denote the survival time. The distribution of T can be characterized by three equivalent functions. The survival function, denoted by S(t), is defined as the probability that an individual survives longer than t:

$$S(t) = Pr(T > t) \quad 0 < t < \infty.$$

here, S(t) is a non increasing function of time *t.*

When survival time is measured as a discrete random variable, T can take on values $t_1, t_2, t_3, \ldots$, with $0 < t_1 < t_2 < t_3 \ldots$, and let the probability function be

$$f(t_j) = Pr(T=t_j) \quad j = 1,2, 3\ldots.$$

The survival function is then

$$S(t) = \sum_{j|t_j \geq t} f(t_j)$$

$$= \sum f(t_j) I_{(t_j \geq t),}$$

where,

$$I_{(t_j \geq t),} = \begin{cases} 0 \text{ if } t_j < t \\ 1 \text{ if } t_j \geq t \end{cases}$$

In this case the hazard function is defined as,

$$h(t_j) = \Pr\,(T = t_j | T \geq t_j)$$

$$= \frac{f(t_j)}{S(t_j)} = \frac{S(t_j) - S(t_{j+1})}{S(t_j)} = 1 - \frac{S(t_{j+1})}{S(t_j)} \quad j= 1, 2, 3, \ldots$$

For continuous variable T, the probability density function of T is

$$f(t)= F'(t) = - S'(t),\ \ t \geq 0,$$

The cumulative distribution function F(t), is defined as the probability that an individual fails before *t:*

$$F(t) = P(T \leq t) = \int_0^t f(x)dx$$

The probability of an individual surviving to time t is given by

$$S(t) = P(T \geq t) = \int_t^{\infty} f(x)dx$$

The hazard function h(t) of survival time T is defined as the probability that an individual leaves the herd or dies given that she was alive just before t (Collett, 2004). Therefore, the hazard function represents the instantaneous culling rate or death rate for an individual surviving to time t.

$$h(t) = \lim_{\Delta t \to 0} \left[\frac{Pr(t \leq T < (t+\Delta t) | T \geq t}{\Delta t}\right] = \frac{f(t)}{S(t)} = \frac{-d\log S(t)}{dt}.$$

The cumulative hazard function is defined as

$$H(t) = \int_0^t h(u)du = -\log(S(t)).$$

Given any one of them, the other two can be derived. For example,

$$S(t) = 1 - F(t) = \exp(-H(t)),$$

$$f(t) = h(t)\exp(-H(t)).$$

Several statistical methods have been proposed for modelling of survival analysis data. Most commonly used methods can be divided into two broad categories: proportional hazard approaches (including the semiparametric Cox model and fully parametric approaches) and accelerated failure time models. These methods have different properties and interpretations, but all may be used to summarize survival data.

1.1.3. Cox Proportional Hazards Model

The Cox proportional hazards model (Cox, 1972) is the most commonly used approach for analysing survival time data. It is a survival analysis regression model, which describes the relation between the event incidence, as expressed by the hazard function and covariates that influence survival time. The Cox model is written as,

$$h(t)=h_0(t) \times \exp\{b_1x_1+b_2x_2+ \ldots +b_px_p\},$$

where the hazard function h(t) is dependent on the number of regressors incorporated in the model, whose impact is measured by the size of the respective coefficients of the covariates (Bradburn et al., 2003). The term h_0 is called the baseline hazard, and is the value of the hazard if all the covariates are equal to zero. The main characteristic of the Cox model is the baseline hazard function is estimated nonparametrically and the survival time is not assumed to follow a particular statistical distribution (Cox, 1972). The covariates then act multiplicatively on the hazard at any point in time, and this provides the key assumption of the PH model: the hazard of the event in any group is a constant

multiple of the baseline hazard. This assumption implies the hazard curves for the groups should be proportional and cannot cross (Lawless, 2003).

The other class of models that have similar approach and interpretation to the previously discussed model are the parametric proportional hazards models (Bradburn et al., 2003). The difference between the two models is that, in the latter, the hazard is assumed to have a specific parametric form when a fully parametric proportional hazards model is fitted to the data, whereas the Cox model imposes no such assumption. However, hazard ratios have the same interpretation in both approaches. A significant effect would mean that the hazard of experiencing the event of interest in a particular class is different from the hazard of experiencing the event in the reference group. For instance, if the hazard ratio is 0.35 this means that class 1 has a 65% lower hazard than the reference group. On the other hand, a hazard ratio of 1.35 means that class 1 has a 35% higher hazard ratio than the reference group (Collett, 2004). Generally a positive coefficient for the covariate means that the hazard is higher, and thus the prediction is worse. A negative coefficient indicates a better prediction for a particular group with higher values of that covariate.

1.1.4. Accelerated Failure Time Model

Accelerated failure time (AFT) model is a parametric model that provides an alternative to the commonly used proportional hazards models such as the Cox proportional hazards model (Newby, 1988). The accelerated failure time model is the logarithm of the failure time as linear function of covariates included in the model (Kalbfleisch and Prentice, 1980). Accelerated failure time models can, therefore, be framed as linear models for the logarithm of the survival time.

$$\log T_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \sigma\varepsilon_i,$$

where β_0 is the intercept, σ is the scale parameter and ε_i a random error term. Differences between the above AFT model and most commonly used linear model are (1) that the random error ε_i non Gaussian, (2) the response is subject to censoring and (3) the dependent variable, T, is the logarithm of time.

When there is no censored record, the ordinary least squares method can be used to obtain point estimates of parameters, where the dependent variable is log T. However in reality, in survival data analysis if there is at least one

censored records, then it is difficult to use the above procedure. However, the maximum likelihood procedure with different distributional assumptions on the random error can be applied for estimation procedure (Lawless, 2003).

For a given distribution of ε, there is a corresponding distribution for T**.** The most common distributional assumptions on ε and the corresponding T distribution can be summarized in the following way. When the distribution of ε is assumed extreme values of one and two parameters the distribution of T would take Exponential and Weibull distribution, respectively. Similarly, when the distribution of ε is assumed logistic, normal and log gamma, the corresponding T distribution will be log-logistic, log normal and gamma distribution (Collett, 2004).

The above different models can also be classified as proportional hazards model (Exponential and Weibull) and proportional odds model (log logistic). The Weibull and Exponential models can be both the accelerated failure time and proportional hazards model. This relationship is summarized in Figure 1.

With regard to interpretation of results, the main difference between the proportional hazards model and accelerated failure time model is that in the former it assumes that the effect of a covariate is to multiply the hazard by some constant, whereas the latter case it assumes that the effect of a covariate is

to multiply the predicted event time by some constant (Kalbfleisch and Prentice, 1980: Collett, 2004; Lawless, 2003).

However, regardless of the similarities and differences between the above mentioned models, the covariates included in the models are assumed to have as a linear effect on the response variable despite the fact that some covariates may have a nonlinear relationship with the response variable.

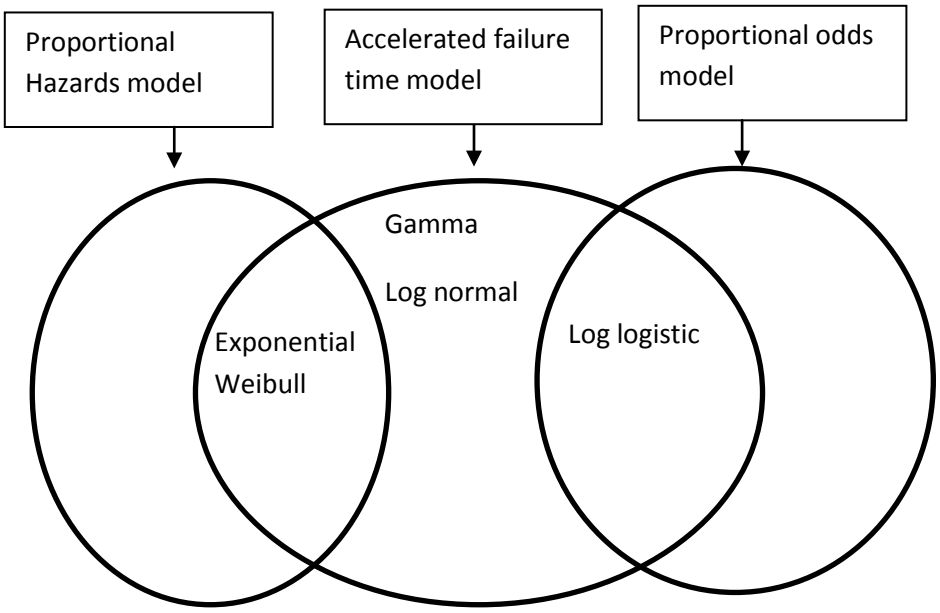


Figure 1. Commonly used AFT models in survival analysis

1.2. Dairy cattle breeding

In dairy cattle breeding the goal is to increase lifetime profit per animal and per unit of time. Profit is a function of production and the time that a cow remains in the herd (commonly called survival or longevity or herd life). Therefore, survival or longevity of cows is a trait of considerable economic importance since it has a significant impact on profitability. Increased longevity is associated with decreased culling and therefore decreased cost of raising or purchasing of replacement females (Allaire and Gibson, 1992; Dekkers, 1994).

However, the statistical analysis of survival data in dairy cattle is not straightforward for several reasons. The distribution of survival time is rarely known and in most cases, extremely skewed; for part of the observations, only a lower bound of survival time is known (for individuals still alive at the end of the study period). Further, the independent variables influencing survival time may themselves vary with time. For instance, current milk production, herd size, disease occurrence and many other factors (Ducrocq, 2002).

Several strategies have been suggested and used to analyze survival data in dairy cattle. These include a simple modeling of a 0-1 variable indicating whether the cow is still alive or dead at any specific time. In this approach, the response variable was considered as a binary trait and analyzed either using a

linear or threshold model (VanRaden and Klaaskate, 1993; Jairath et al., 1998; Vollema and Groen, 1998; Boettcher et al., 1999; Sewalem et al 2007). Typically such type of data has a skewed distribution and analysis using traditional linear models may not be appropriate. Survival analysis using a proportional hazards model as suggested by (Smith and Quaas, 1984) is an alternative method for animal breeding survival data. Ducrocq et al. (1988) showed that proportional hazard models could be used for the analysis of survival data in dairy cattle. Ducrocq and Solkner (1998) developed the Survival Kit typically used by animal breeders for large populations using a Weibull model (Schneider et al., 2000; Vollema et al., 2000; Vukasinovic et al., 2001; Ducrocq, 2002; Sewalem et al., 2004) where several covariates are fitted as a linear effect.

Additionally, in dairy cattle production there are several environmental factors (covariates) that influence the survival of cows and those factors need to be accounted in the model in order to get reliable estimates. In this regard Sewalem et al. (2005) studied survival of Canadian dairy cows using censored linear regression model that included several covariates. In those analyses some of the continuous covariates are grouped and fitted to the model as a class effect. This grouping of covariates may also result in loss of information. In addition, some covariates may have nonlinear effects on the response variable.

In this case, the traditional linear survival models may not be the right approach for such type of analysis.

1.3. Regression analysis

In studying the relationship between a response and a set of independent variables, the mean response variable is often assumed to be a linear regression function of the regressors. In many practical situations, particularly in biological studies, however, the linear model is not complex enough to capture the underlying relationship between the response variable and its associated covariates. Some components can be highly nonlinear. A natural generalization of the linear model is to allow only some of the predictors to be modeled linearly, with others being modeled nonlinearly (Carroll et al., 1997).

By nature, biological organisms are usually characterized by their complexity. A physiological response could be either a discrete variable represented by two classes such as diseased or healthy or continuous variable such as body weight, height and average daily gain over a specified period of time. These traits are the result of a composite system of biological and chemical reactions involving an organism. Often, a linear regression analysis is

commonly used to understand the relationship between a response variable and its explanatory variables. In this regard, usually, parametric models are used to analyze the data. In most applications, however, any parametric model is at best an approximation to the true model (Box, 1987) and looking for a suitable model is not easy.

For several years, to overcome the intricacy caused by the high dimensional data and to develop a flexible model that handles both a linear and nonlinear effects of the explanatory variables, several models have been developed using nonparametric or semiparametric regression models. Hardle and Stoker (1989), Powell et al. (1989) and Newey and Stoker (1997) investigated single-index models. Further, Carroll et al. (1997) and Xia and Stoker (2006) extended the single-index model to the generalized partially linear single-index model. Even so, these models are used to study the relationship between the response and the predictor variables when data are fully observable.

In practice, however, survival data are often subject to censoring. When it occurs, the incompleteness of the observed data may induce a substantial bias in the sample. Several approaches have been developed to overcome the associated difficulties in some specific models, including the partial likelihood method in the Cox proportional hazards model. Several studies were carried out

using parametric and semi-parametric censored regression models by assuming a parametric regression form or assumed that the error distribution is parametric (Buckley and James, 1979; Koul et al., 1981; Lai et al., 1995; Powell, 1986; Duncan, 1986; Fernandez, 1986; Horowitz, 1988; Ichimura, 1998; Lewbel, 1998; Buchinsky and Hahn, 1998; Heuchenne and Van Keilegom, 2007). The Cox (1972) regression model plays a central role in survival analysis. However, the Cox model has its limitations in dealing with more sophisticated covariate effects arising from real data. Several studies have tried to extend the Cox model to include nonparametric or semi-parametric covariate effects on censored failure data; for example, Dabrowska (1987) and Nielsen and Linton (1995), where the hazard function is completely unspecified. Fan et al. (1997) relaxed the fully nonparametric specification.

Fan and Gijbels (1994) proposed a censored nonparametric regression estimator based on using only a univariate regressor. Wang and Zheng (1997) and Liang and Zhou (1998), however, extended the univariate regressors to multiple regressors. Further, Singh and Lu (2002) studied censored nonparametric additive regression models based on some special data transformations. Lu et al. (2006) examined a class of partially linear single-index proportional hazards models for survival data. Lewbel and Linton (2002) and Chen et al. (2005) considered identification and estimation of a

nonparametric location-scale model under fixed censoring. However, in most biological and agricultural fields censoring is random hence the application of these models is limited.

In a simulation study, Lu and Cheng (2007) investigated a class of partially linear single-index models under random censoring using the accelerated failure-time model without the specification of the distribution function of the response variable. In their study only two traits were included in the parametric and nonparametric component of the single index model.

1.4. Rationale for the thesis

Longevity is a highly desirable trait that considerably has an effect on overall profitability of the dairy industry. To this effect the use of an appropriate statistical model is essential in the analysis of survival data. To date, in the case of survival analysis in dairy cattle, the factors that influence the survival of cows are often fitted in the model as linear multiplicative effects although several factors are believed to have nonlinear effects. Further, most of the continuous effects that influence the survival of cows are grouped together and fitted in the model as a class effect which may result in a loss of information by doing so. Identifying those covariates which have linear or nonlinear effects on survival of cows would be helpful in extending the scope of survival analysis in dairy science. This will ultimately lead an improvement in the methodology resulting in enhanced profitability for the dairy industry.

1.5. Objectives of the thesis

The overall objective of this thesis was to explore the association of the response variable and various regressors in dairy cattle breeding data using the various survival models in general and the partially linear single-index survival model via the accelerated failure time model in particular. This model has not previously been applied to real data either in dairy science or any other area.

Specific objectives are:

1. To explore aspects of nonlinearity not captured by traditional models used in dairy science,
2. Explore the use of single-index in reducing the dimensionality in dairy science,
3. Extend the model developed by Lu and Cheng (2007) to accommodate more than two traits in each component of the model (parametric and nonparametric) and to estimate parameters of interest using real data sets,

4. The other important aspect of this thesis is the calculation of standard errors of parameters of PLSISM via bootstrap approach and their application to real data. Analytical expressions for standard errors are complex and intractable and the use of bootstrap avoids these difficulties. This is the first application of this technique to obtain standard errors.

Chapter 2

2.1. Materials and Methods

2.1.1. Description of the data sets

In this study two sets of data, namely calf survival data and culling data were used which are described in detail in the following sections.

2.1.1.1 The calf survival data

Calf mortality is a serious problem in dairy cattle production. Apart from raising labour and veterinary expenses, it mainly increases replacement cost. Therefore, identifying calves that have a good genetic potential for survival is an important aspect of dairy farming.

Calf survival data was obtained from a commercial rearing facility located in New York, USA (Henderson, 2011). The center raises heifers to varying target ages based on the contractual arrangement between the facility and the farm of origin or producers (Henderson, 2011). Most of the calves

typically arrive at the facility in the first 2-3 days of life. Upon arrival, calves are weighed, measured, and identified with a unique identification number. Moreover, calves are sampled for serum total protein levels to provide a basis for calf survival contractual warranties. Calves were reared in individual pens in barns of 48 calves, and were weaned at approximately 7 weeks of age. For each heifer, information such as; source farm of origin, calving ease, birth date, arrival date, arrival weight and height, serum total protein, disease treatments during the preweaning period, weaning date, weaning weight and height, death and culling occurrences were recorded. Detailed description of the data set can be found (Henderson, 2011).

Data were from December 1998 to June 2008 and were obtained from the participating heifer rearing facility. Records with missing identification, non-identified source farm of origin, calves which arrived older than one week of age, calves with incorrect weaning dates, and duplicate records and some mismatch records were excluded from the analysis.

Some of the fixed continuous variables were grouped to fit as a class effect in the model. These include arrival weight class (defined as 1 = 49-84 lbs; 2 = 85-92 lbs; 3 = 93 lbs and above), weaning weight class (defined as 1 = 77-136 lbs; 2 = 137-153 lbs; 3 = 154 lbs and above), total protein class (defined as 1 = 30-56 g/L; 2 = 57-64 g/L; 3 = 65 g/L and above), season of birth class

(seasons were July to September, October to December, January to March and April to June), calving ease score (defined as 1 = unobserved/unassisted, 2 = easy pull, 3 = hard pull, excessive force or surgery needed), disease incidence class (defined as 0 = no disease experienced, 1 = one disease occurrence, 2 = two disease occurrences, 3 = three or greater disease occurrences) and finally source farm of origin or herd.

2.1.1.2. The culling data

In dairy cattle production, culling is the term used to describe the disposal of cows from a dairy herd for various reasons such as low production, impaired reproductive performance and less resistance for various diseases compared to their herd mate averages. The process of culling provides the producer with the opportunity to improve genetic progress and thereby improve productivity and profitability.

Culled cows represent a substantial loss to the dairy industry because of increased raising of replacement heifers and other associated costs such as veterinary cost. The culling procedure can have a considerable effect on the whole dairy venture. In other words, it can influence the genetic improvement

program, the health management, the reproductive performance and the overall investment policy for the stakeholders in the industry.

In dairy production the producers undertake intense culling of cows between the first and second lactation periods; thereafter the culling procedure slows down. Therefore, identifying the specific risk factors that are associated with culling dairy cows is crucial during this time frame. From the breeders perspective, identifying those risk factors that greatly influence culling procedure and understanding the relationships between these factors are of vital importance to develop the correct model for evaluation purpose.

For this analysis, data were obtained from lactation and type classification records extracted for the Canadian May 2011 genetic evaluation of the Holstein breed (CDN, 2011). Length of productive life time was defined as time from first calving to second calving, death, or culling. Censored records represented cows being sold for dairy purposes, exported or leased to another herd or cows still in the herd. A lifetime record was considered to be completed (uncensored) if the cow received a termination code, indicating that the cow was removed for any reason from herd. Records associated with missing identification, incorrect calving dates and age at first calving outside 18-40 months range were excluded from the analysis. Type information consisted of phenotypic type scores of body condition score a descriptive trait evaluated on ordinal scale 1 to 9. Body

Condition Scoring (BCS) is a subjective system of evaluating a cow's level of body condition (amount of stored fat) and assessing a numeric score to facilitate comparisons between dairy cows. This provides an indication of the energy status of dairy cattle. Essentially, body condition scoring provides an objective indication of the amount of fat cover on a dairy cow. This evaluation is accomplished by assigning a score to the amount of fat observed on several skeletal parts of the cow. Cows receiving a score of 1 were considered as thin and progressively those cows receiving a score of 9 were considered as fat.

For this data set the covariates included in the model were as follows: effect of herd-year-and season of calving with year of calving from 2005 to 2010 and seasons of calving were January-March, April-June, July-September and October-December); effect of the annual change in herd size with three classes (decreasing=for a decrease in herd size of <-5%, nearly unchanged=no appreciable change >= -5% to <=10% and increasing=for increasing in herd size of >10%); effect of the type of milk recording supervision with two classes (0=unsupervised and 1=supervised); effect of age at first calving in months; effects of milk, fat and protein yields. The latter effects were calculated as within herd-year deviations with three classes for each, low = cows producing less than 0.3 standard deviations below the herd-year average, average = cows producing between 0.3 standard deviations below and 0.5 standard deviations

above the herd-year average and high = cows producing above 0.5 standard deviations of the herd-year average and body condition score.

After editing, the numbers of records included in the two data sets were 1,804 and 1,744 for calf survival and culling data, respectively. The data were summarized using numerical and graphical presentation of the survival time using the Kaplan-Meier (1958) estimate of survival function.

The primary goal of the study is to assess which of the covariates were useful in predicting mortality of calves from birth to exit and survival of cows from first calving to second calving. As a preliminary analysis, the Cox proportional hazards model was used to identify the covariates that have significant effect on survival time. Test of the proportionality assumption and goodness of fit (Collett, 2004) were performed for the two data sets. Further, the different accelerated failure time models were implemented to analyze the data (Exponential, Weibull, Lognormal and Logistic) and the four models were compared using the Akaike information criteria (AIC),

$$AIC= -2logL + 2p,$$

where logL = the log likelihood of the model, p = the number of model parameters. Based on the AIC the best model was selected to investigate the relationship of each covariate with the survival time for the two data sets. The

Weibull model was the better fit for the two data sets. In addition to the ordinary Weibull model the following partially linear single-index survival model was used to analyze the data (Lu and Cheng, 2007),

$$Y = \beta_0^T V + \lambda_0(\alpha_0^T X) + \sigma(V,X)\epsilon \quad \text{with } \|\alpha_0\| = 1,$$

where Y is the log survival time, V and X are the associated regressors, q and p vectors, respectively. For ease of understanding the covariates were denoted using two different symbols: V and X which comprise the parametric and nonparametric component, respectively. The parametric component is characterized by an unknown q vector with parameter β_0. The nonparametric component is characterized by λ_0, an unknown smooth univariate function defined on the real line, and an unknown projection p-vector parameter α_0. $\sigma(\cdot,\cdot)$ is the conditional variance representing possible heteroscedacity; $\|.\|$ denotes the Euclidean norm. The constraint $\|\alpha_0\| = 1$ on the single-index coefficient parameters is required for parameter identifiability. Assume that (V, X) and ε are independent, E(ε) = 0 and var(ε) =1. Let C be the random censoring time associated with the log survival time Y. Assume C is independent of (V, X, Y). Denote Z = min(Y, C) and $\delta = I (Y \leq C)$. The

observations are $\{(V_i, X_i, Z_i, \delta_i) : i = 1, \ldots, n\}$ which are regarded as a random sample from the population (V, X, Z, δ).

2.3.4. Estimation procedure

Since the distribution of error in the model is not specified, applying the full likelihood function in the above model is not possible. Therefore, a quasi–likelihood estimation procedure was implemented using an iterative minimization algorithm (Lu and Cheng, 2007). The term quasi-likelihood here is similar to that of Wedderburn (1974) in that only first and second order assumptions are made about the distribution of the response Y.

Let θ = (α, β) be the vector of model parameters and if the data is fully observed, i.e., Z≡Y, the quasi likelihood estimator of $\theta_0 = (\alpha_0, \beta_0)$ and λ_0 are the minimizers of the following quasi-likelihood function of $\{(V_i, X_i, Z_i, \delta_i) : i = 1, \ldots, n\}$,

$$\ell_n(\theta, \lambda) = \sum_{i=1}^{n} [Y_i - \{\beta^T V_i + \lambda(\alpha^T X_i)\}]^2 \text{ with } \|\alpha\| = 1$$

This model is similar to the generalized linear single-index models as presented by Carroll et al (1997) for complete data. This procedure encounters difficulties in estimation due to censoring and the involvement of the nonparametric function λ.

To overcome this difficulty first synthetic data or pseudo response were produced using: $Z_{i\widehat{G}} = (1+ \phi) L_{i\widehat{G}} - \phi K_{i\widehat{G}}$ (following the procedure of Lu and Cheng, 2007). Here $L_{i\widehat{G}} = \int_{-\infty}^{\infty}(\frac{I[Z_i \geq s]}{(1-\widehat{G}(s-))} - I[s < 0])\, ds$, $K_{i\widehat{G}} = \frac{Z_i \delta_i}{(1-\widehat{G}(Z_i-))}$, ϕ is a tuning parameter which control the weights put on censored and uncensored observations and I(.) is the indicator function. $(1-\widehat{G}(.-))$ is the left continuous version of Kaplan-Meier estimator defined by

$$1-\widehat{G}(t) = \prod_{k=1}^{n}\left[\frac{n-i}{n-i+1}\right]^{I[Z_{(i)} \leq t, \delta_{(i)}=0]},$$

$Z_{(1)} \leq Z_{(2)} \ldots Z_{(n)}$ are the order statistics of Z-sample and δ_i is the associated δ with $Z_{(i)}, 1 = 1, 2, \ldots, n$. The observed data $(V_i X_i Z_i, \delta_i)$ is replaced by $(V_i, X_i, Z_{i\widehat{G}})$. The pseudo responses are such that when G is known, the expected value of Z_{iG} equals the expected value of Y, i.e. $E(Z_{iG}) = E(Y)$. Thus the censored observations are unbiasedly transformed to pseudo responses, which approximate or impute the unobserved values. When G is unknown, the Kaplan-Meier estimator $\widehat{G}$ was substituted for G. This class of transformation was introduced by Fan and Gijbels (1994), and Koul et al. (1981). Using the transformed data, both parametric (β_0) and nonparametric (λ_0) were estimated by applying the local linear fit to the quasi log likelihood iteratively.

Two well-known qualities of the local linear fit are the reduction of the bias for the estimation of the nonparametric function and the avoidance of

boundary effects. Suppose that $\lambda(\cdot)$ is continuously differentiable. Then in a neighbourhood of a fixed point u, we can write $\lambda(v) \approx a_0 + a_1 (v - u)$, where $a_0 = \lambda (u)$ and $a_1 = \lambda'(u)$. This is called the local linear fit.

Let $W(\cdot)$ be a kernel with a given bandwidth b and a given parameter vector θ. One can obtain local estimators $\hat{a}_0 \equiv \hat{a}_0(u: b\, \theta)$, $\hat{a}_1 \equiv \hat{a}_1(u: b\, \theta)$ by minimizing the following local quasi log likelihood

$$\ell_n(a_0, a_1)\,) = \sum_{i=1}^{n} [\, Z_{i\hat{G}} - \{\beta^T V_i + a_0 + a_1(\alpha^T X_i - u)\}]^2\, W_b\, (\alpha^T X_i - u),$$

where $W_b(.) = b^{-1}W(.|b)$, and u is a fixed real number.

When the true parameter vector θ is unknown, in order to obtain estimators for the model, we need to iteratively update the estimates of the nonparametric component $\lambda_0(.)$ and the parametric components $\theta_0 = (\alpha_0, \beta_0)$. The iterative algorithm consists of the following steps,

Step 1: Treat the pseudo-responses $Z_{i\hat{G}}$ as complete data and apply the estimation procedure for the partially linear single-index models, to obtain initial estimates $\hat{\propto}$ and $\hat{\beta}$ of α_0 and β_0, respectively, with the restriction $\|\hat{\propto}\| = 1$ and $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$.

Step 2: Find $\hat{\lambda}(u: b, \hat{\theta}) = \hat{a}_0$ as a function of u by maximizing the local quasi log-likelihood with respect to a_0 and a_1 with fixed $\theta = \hat{\theta}$ and a suitable bandwidth b

Step 3: Update $\hat{\theta}$ by minimizing the following equation with respect to $\theta = (\alpha, \beta)$

$$\sum_{i=1}^{n} [\, Z_{i\hat{G}} - \{\beta^T V_i + \hat{\lambda}(\propto^T X_i : b, \hat{\theta})\}]^2$$

Step 4; Cycle steps 2 and 3 until convergence is achieved.

Detailed description of the estimation procedure can be found (Lu and Cheng, 2007).

The software that handles the above procedures was kindly provided by Lu and Cheng (2007). However, the software does not calculate the standard errors of the parameters. Therefore, the bootstrap approach was used to calculate the standard error of estimates. Five hundred independent bootstrap samples with replacement were used. For each independent sample drawn, the above aforementioned model was fitted and the corresponding parameters were estimated along with the nonparametric component as described in the above estimation procedure (for pseudo R code see Appendix A).

Chapter 3

3.1 Results

3.1.1. Descriptive statistics

Some descriptive statistics such as the total number of records, the number of censored and event observations, the average, minimum and maximum time in days for the censored and failure observations for the two data sets (culling and calf survival data) are presented in Table 1. As shown in Table 1, the percent censored and uncensored records for the culling data set are 63.8 and 36.2, respectively. The corresponding figures for calf survival data are 90.8 and 9.2. The censoring proportion for the calf survival data is higher than the culling data set.

Table 1. Descriptive statistics for the culling data and calf survival data

	Culling data		Calf survival data	
	Censored	Failure	Censored	Failure
Number of records	1113	631	1638	166
Minimum time (days)	6	1	800	46
Average time (days)	318	208	802	768
Maximum time (days)	600	425	807	365

3.1.2. Survival Distribution

Survival time distributions for the two data sets were estimated using the Kaplan-Meier (KM) methods and results are presented in Figure 2. The Kaplan-Meier survival time distribution shows that neither data set reached the median survival time.

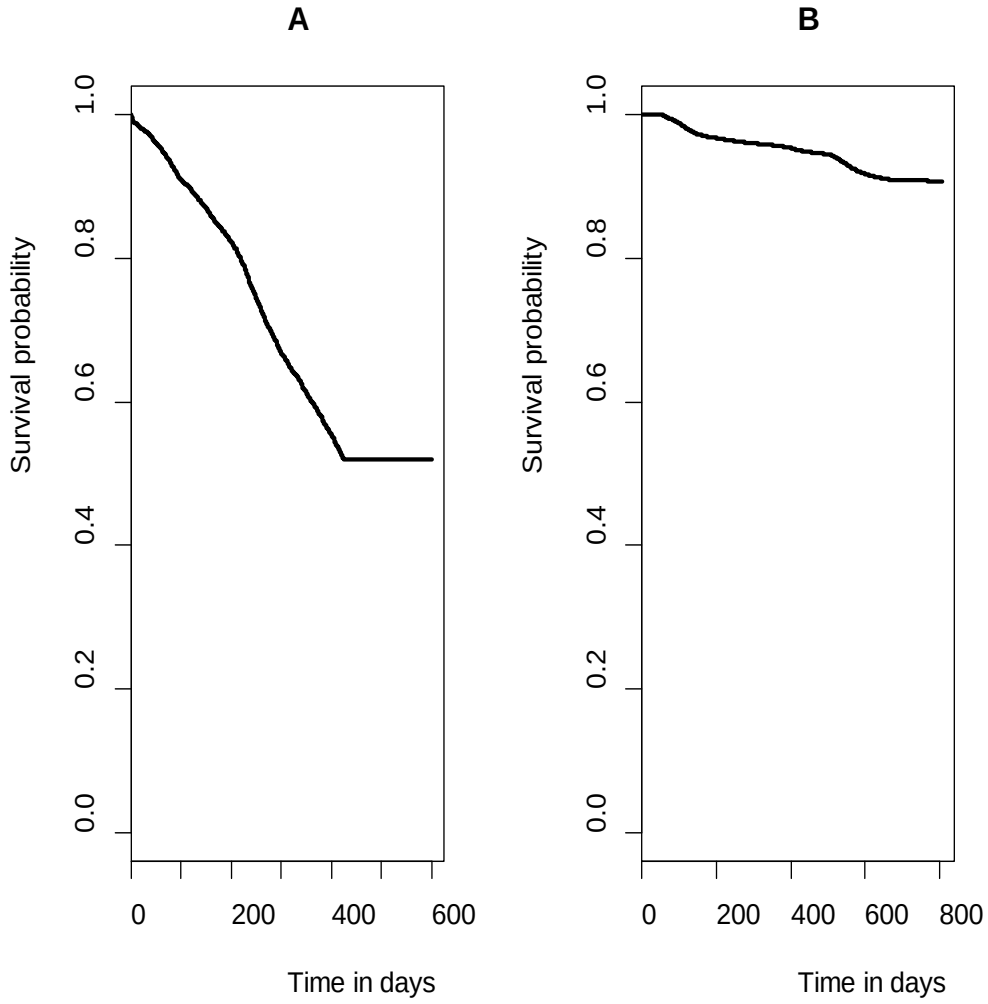


Figure 2. Kaplan-Meier estimates for (A) the culling data and (B) calf survival data.

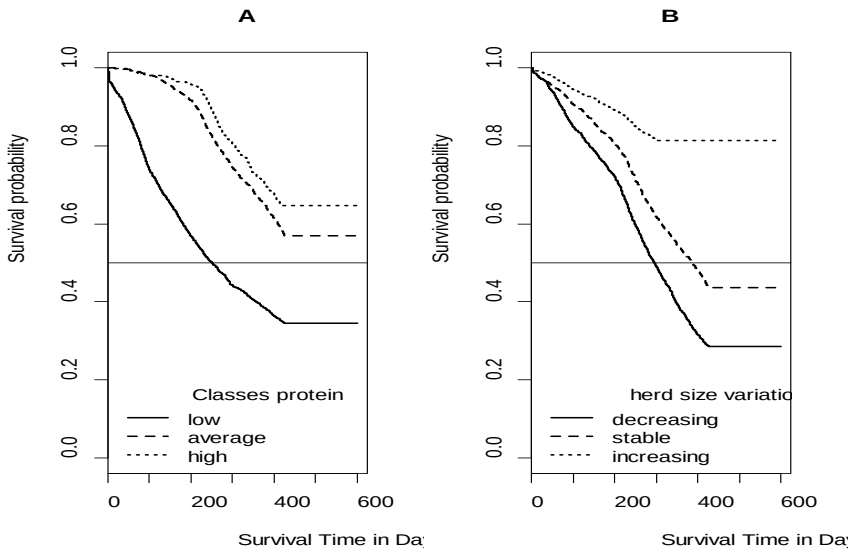


Figure 3. Kaplan-Meier estimates for the culling data for classes of protein production and herd size variation.

The survival time distribution for the three classes of protein group and herd size variation was estimated using the Kaplan-Meier methods for the culling data set. Figure 3A shows that cows that produce lower protein production than the other groups had lower survival distribution than the other groups. Figure 3A shows a clear difference among the three classes of protein group in the KM survival time distribution. Cows that produce low protein production have lower survival distribution than cow in the other groups. The percentage of culled cows was 66.79, 28.04 and 5.18% for low, average and

high protein classes, respectively. The average failure time was 149.39, 265.15 and 292.50 days for low, average and high protein classes, respectively. The median survival time for the low protein class was 140 days. The log-rank test was used to test the survival time distribution among the three protein classes and the result showed significant differences ($P < 0.01$) among the three classes.

Figure 3B also shows the estimated survival distributions of the three classes of annual herd size variation. Herds with decreasing trends in annual herd size had lower survival distribution than herds that had stable herd size. Herds with increased annual herd size variation had the highest survival distribution compared to the other groups. The log rank test showed that the three classes of herd size variations as well as the protein production had significant differences ($P < 0.01$) in survival time distribution.

Figure 4 shows the survival time distribution for the calf survival data for the weaning weight and average daily gain group. For this particular analysis weaning weight group was classified as low when the weaning weight was below 130 lbs and high otherwise.

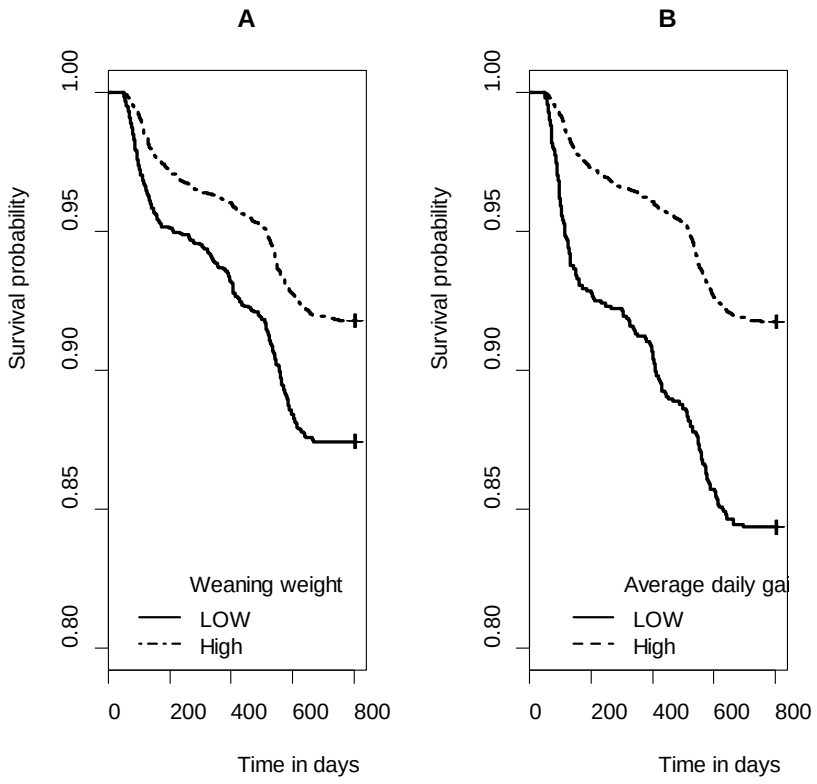


Figure 4. Kaplan-Meier estimates for the calf survival data for weaning weight (A) and average daily gain (B)

3.1.3. Cox Proportional Hazards Model

In order to comprehend the nature of association between the survival time and the various regressors influencing the two data sets, i.e., the culling data and calf survival data, the Cox proportional hazards model was employed. The Cox model is one of the most commonly used semiparametric models in survival analysis. This analysis was also used to identify which variable has significant effect on survival time. The summary of results of these analyses is presented in the following section for the culling and calf survival data.

3.1.3.1. Culling data

Preliminary analysis of this data set showed that all effects included in the model had significant effect ($P < 0.01$). The preliminary analysis also showed that there was no significant interaction among the main effects. However, herd, year and season were included as the effect of herd-year-season. The quadratic effect of level of fat production was significant ($P < 0.01$) and, therefore, it was included in the final model. For the culling data the final model of main effects included: herd-year-season, herd size variation, supervision, production class (fat, protein, milk and quadratic effect of fat production), age at first calving, and body condition score. In this study, the

effect of herd-year-season was fitted in the model as fixed as well as random effect. The herd-year-season variance estimated was 1.14. The parameter estimates, standard errors along with the P-values are presented in Table 2.

Table 2 shows that all effects included in the model had a significant effect ($P < 0.01$) on survival of cows to the second calving, whereas age at first calving (AFC) had significant effect ($P < 0.05$).

Table 2. Summary of parameter estimates for the culling data

	Estimate	SE	z	P
HYS	-0.0002033	0.00006845	-2.969	0.001
HSV	-0.62440	0.023382	-26.702	0.001
SUP	0.291510	0.04112	7.087	0.001
MILK	0.000335	0.00000985	34.012	0.001
PROT	-0.016350	0.0009041	-18.084	0.001
FAT	0.002824	0.0008514	3.316	0.002
FAT2	0.0000315	0.0000024	13.057	0.001
AFC	-0.046630	0.0218	-2.1312	0.047
BCS	-0.179300	0.01153	-15.549	0.001

HYS = herd-year-season; HSV= herd size variation; SUP = herd supervision; MILK=milk production; PROT = protein production; FAT =fat production; FAT2=fat production square; AFC=age at first calving; BCS = body condition score.

3.1.3.2. Calf survival data

Preliminary analysis of the calf survival data also indicated that the variables that affected calf survival were herd-year-season, calving ease score, arrival weight associated with birth weight, weaning weight, total protein, average daily gain and the number of disease incidence. The parameter estimates for this data set along with the standard errors, z and P-values are presented in Table 3.

Table 3 shows herd-year-season, arrival weight, weaning weight, average daily gain (linear and quadratic) and number of disease incidence, calving ease score had highly significant ($P < 0.001$) effects on calf survival. However, the quadratic effect of arrival weight was significant ($P < 0.05$). In a separate analysis, herd year-season was fitted as a random effect and the variance of this effect was 0.54.

Table 3. Summary of parameter estimates for the calf survival data

	Estimate	SE	z	P
HYS	-0.0025096	0.00039	-6.422	0.001
AWTG	-0.0160656	0.00387	-4.511	0.001
CE	0.084047	0.01035	8.12	0.001
WWG	-0.0148076	0.0021791	-6.795	0.001
ADG	-0.6018650	0.066381	-9.067	0.001
TPG	0.1925039	0.0737718	2.609	0.001
ADG2	0.1062786	0.0059467	17.872	0.001
AWTG2	0.0001704	0.0001182	1.441	0.040
NDI	0.1852178	0.0438341	4.225	0.001

HYS = herd-year-season; AWTG = arrival weight; WWG = weaning weight; ADG = average daily gain; TPG = total protein; NDI = number of disease incidence and CE =Calving ease score; ADG2 = average daily gain square; AWTG2 = arrival weight square

A detailed analysis of the effect of categorical variables on survival of cows was carried out. In order to do this, some of the continuous covariates were grouped and fitted as a class effect as described in the previous section. The results were expressed as hazard ratio, defined as the ratio between estimated risk of being culled under the influence of certain environmental factors and the average risk (or reference risk), which is usually set to one. Values larger than one indicate higher culling risk associated with that environmental factor. Hazard ratio smaller than one indicate lower culling risks (i.e., increasing effect of environmental factor on longevity). Hence, the annual change in herd size was associated with relatively higher risk of culling in shrinking herds compared to stable herds. Those herds with annual increase in size had also lower culling rates than the stable herds. The culling risk associated with the type of milk recording supervision shows that cows in unsupervised herds had 1.12 times higher risk of being culled than supervised herds.

The effect of within herd-year production deviations (milk, fat and protein) had significant influence on the culling rate. The hazard ratio for cows producing 0.3 standard deviations below the herd-year mean had higher risk of being culled than average producers for milk. High producing cows for milk are less likely to be culled compared to the average producers. The influence of

within herd-year protein yield deviations follows the same trend as that of milk yield. The hazard ratio for cows producing 0.3 standard deviations below the herd-year-mean has a higher risk of being culled than do average producers for protein. With regard to fat production, cows producing 0.3 standard deviation below the herd mean are more likely to be culled compared to the average producing cows. However, unlike protein and milk production cows producing above the herd year average for fat yield had slightly higher risk of being culled compared to the average producing cows.

The effect of age at first calving had also significant influence on the survival of cows to the second calving (Figure 5). The risk of being culled was higher for older heifers than heifers calving at an age between 24 and 28 months. Late calvings are usually associated with herd management, fertility or other health problems and these factors are likely to increase the risk of culling. Moreover, cows with delayed calvings are less profitable due to higher rearing costs. Figure 2 also shows a trend towards a higher risk of culling for cows first calving at less than 21 months of age presumably younger cows are at a greater risk of having calving difficulties or dystocia.

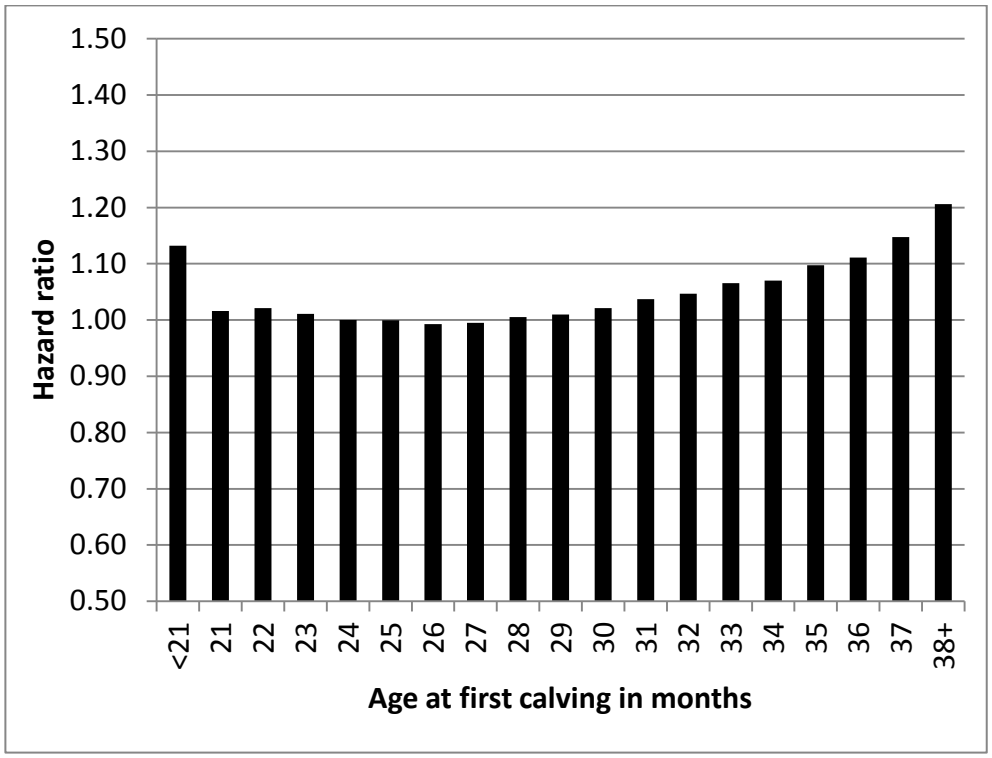


Figure 5. Hazard ratio by class of age at first calving (relative risk of culling rate for AFC at 24 months was set to 1).

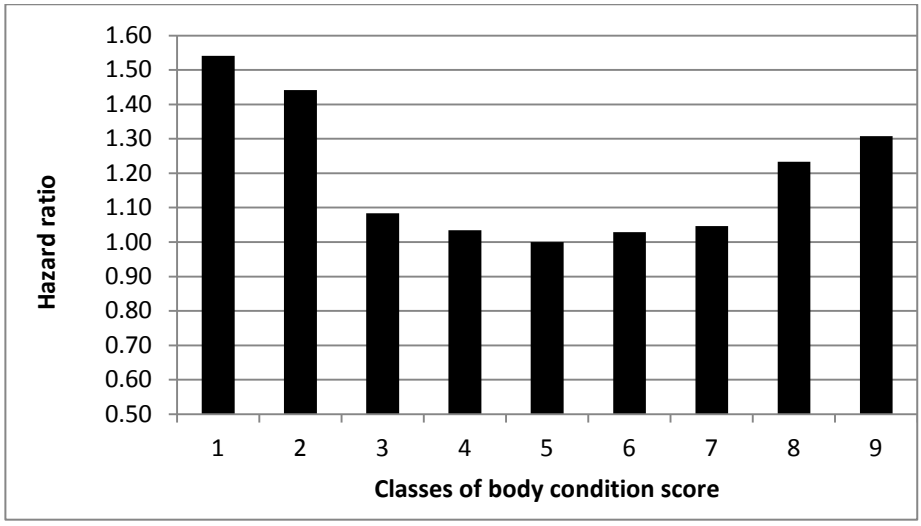


Figure 6. Hazard ratio by class of body condition score (relative risk of culling rate for score 5 was set to 1).

Figure 6 shows a clear nonlinear relationship between body condition score and survival of cows to the second calving. Cows with score of 1 (classified as lean) are 54% more likely to be culled compared to the reference class (score=1). Similarly, cows with score of 9 (classified as fat) are less likely to stay longer in the herd compared to the reference group. Overall, Figure 6 shows that body condition score is an intermediate optimum trait indicating that neither fat nor lean cows are desired for breeding purposes.

From the above analyses, one can infer that the relationship between age at first calving (Figure 5) and body condition score (Figure 6) with the survival of cows to the second calving is somewhat nonlinear and the application of parametric linear survival models may not be the right approach for this data set. Furthermore, the relationship between fat production and survival of cows is indicative of nonlinearity, where low and high producing cows tended to have higher risk of being culled compared to the average producing cows.

A closer look at the effect of each categorical variable on calf survival indicated significance differences within each categorical variable on calf survival. For instance, for calving ease group, calves born with easy pull, hard pull or surgery were at higher risk of dying contrasted to the unassisted calving. The hazard ratios of easy pull, hard pull or surgery were 1.08 and 1.20, respectively compared to unassisted calving (Figure 7). The influence of calving ease score on survival of calves might imply that as calving difficulty increases, calves experience distress and causes physical trauma during parturition which in due course influences the survival of calves. Difficult births have a remarkable effect on calf survival and health. When cows have to be assisted or have surgery during birth, there are often lasting effects on the calf. Calves may suffer from anoxia, lack of oxygen and may have damage to joints, bones or organs. Consequently, the calf feels weak and is slow to stand

or nurse the cow. As a result many calves suffer from failure of passive transfer and are more susceptible to disease.

Total protein group also affected risk of mortality of calves. As total protein group score increased, risk of mortality of calves decreased. The hazard ratios for total protein group 1, 2, 3 and 4 were 1.08,1.02, 1.00 and 0.96, respectively.

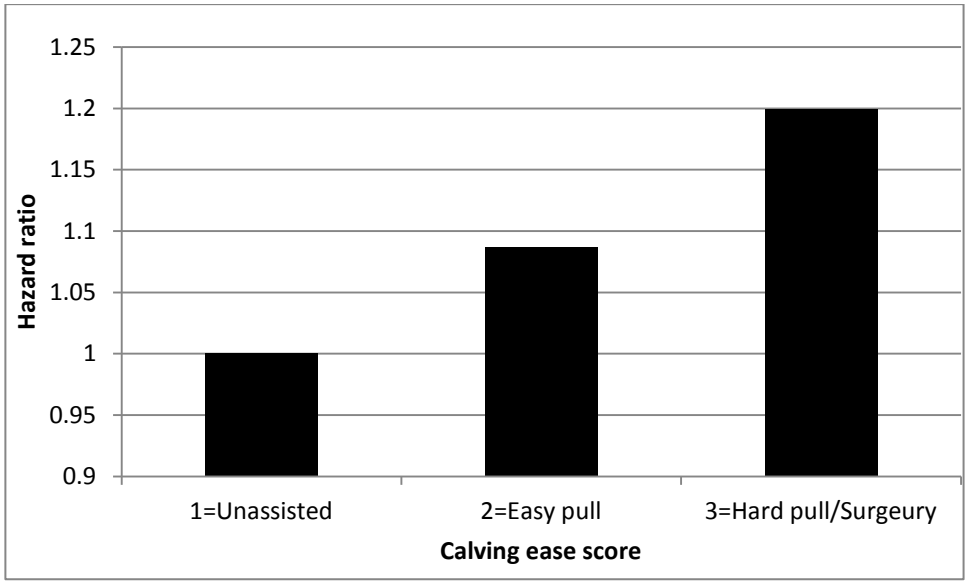


Figure 7. Hazard ratio of calf mortality for calving ease score

Weaning weight of calves influenced survival of calves significantly ($P < 0.001$). Figure 8 exhibits the hazard ratio for weaning weight classes. The hazard ratio of heifers who weighed between 75-111 lbs at weaning was 2.98 times greater than a heifer who weaned at an average weight of 131-150 lbs. On the other hand, heifers who weaned in the heaviest weaning weight class (>175 lbs) were approximately 60% more likely to survive than heifers who weaned in the average weaning weight class. Heifers with heavier weaning weights are more likely to survive to maturity than heifers with average or below average weaning weights. Increased weaning weights could be associated with a combination of less disease experienced throughout the preweaning period, as well as having genes for increased growth and general disease resistance.

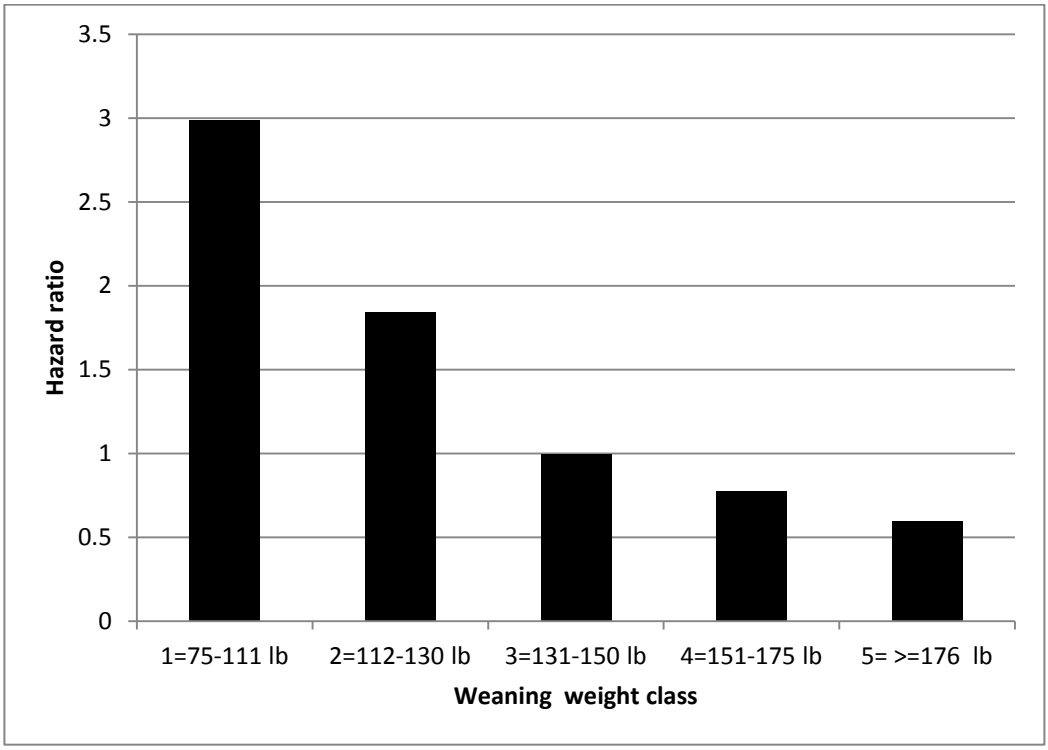


Figure 8. Hazard ratio of calf mortality for weaning weight class

Arrival weight also influenced the risk of mortality of calves during the study period. As revealed in Figure 9 calves were more likely to die with low arrival weight group compared to the average arrival weight group. Similarly, higher arrival weight group were also found at higher risk of dying compared to the average group. The higher risk of dying with higher birth weight might be associated with calving difficulty and as a result, have more health issues in

early life, increasing the risk of mortality. With regard to the arrival weight (birth weight), as presented in Figure 9 there appears to be an optimal weight class (84-92 lbs) in relation to the survival of calves.

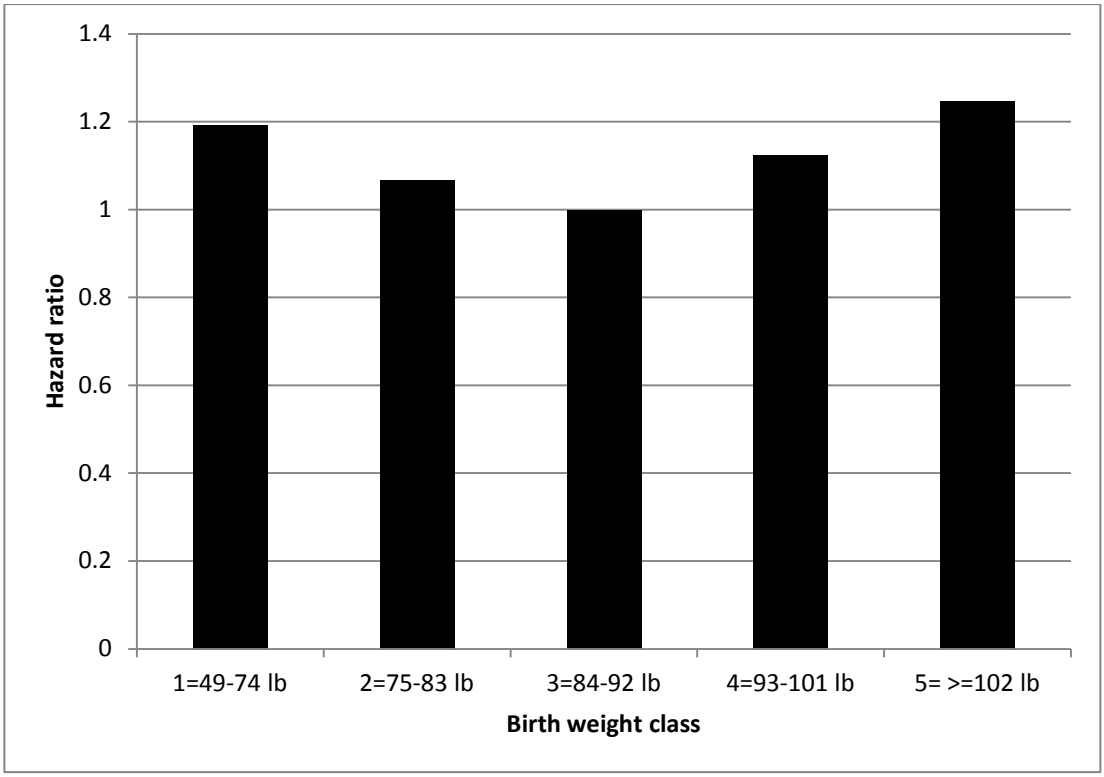


Figure 9. Hazard ratio of calf mortality for arrival weight class

Average daily gain also influenced the survival of calves till exit. For instance calves that had low average daily gain were at 1.44 times risk of dying compared to those calves with high average daily gain.

The relationship between the number of disease incidences and survival of a calf shows that calves who had disease two and three times had a considerably increased relative risk of dying compared to calves that had no disease at all. For instance, calves experiencing 2 or 3 or more disease occurrences requiring treatment prior to weaning were 1.16 and 1.63 times more likely to die prior to exit than calves that experience no preweaning disease.

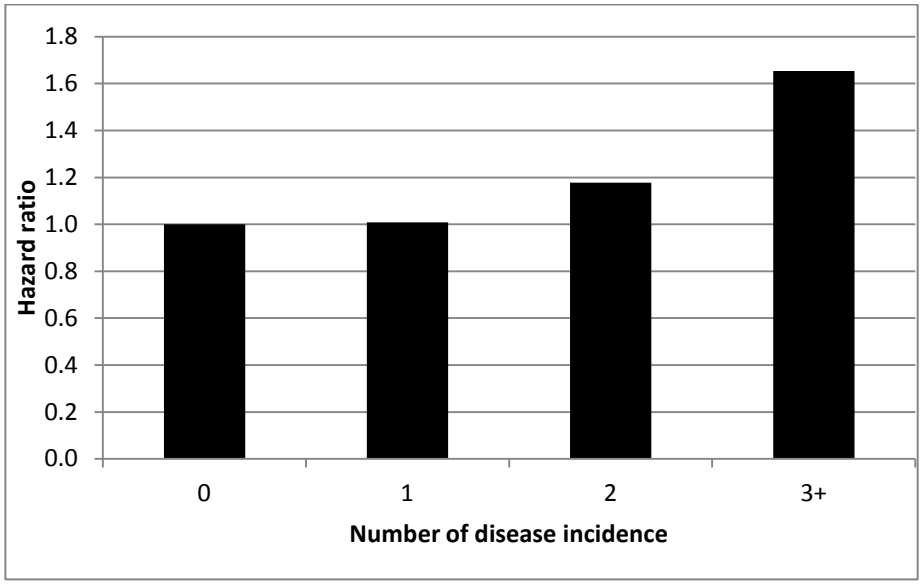


Figure 10. Hazard ratio of calf mortality for number of disease incidences

Figure 10 also shows that there is no noticeable difference between calves that experienced no disease and calves that experienced one disease before exit.

After fitting the Cox proportional hazards model, the assumption of proportional hazards and the goodness of fit were assessed for each data set. The proportional hazards assumption was checked by testing a non-zero slope in a generalized linear regression of the scaled Schoenfeld residuals on functions of time. When the correlation of the scaled Schoenfeld residuals with the time variable is significant then it is an indication of a violation of the proportional hazard assumption. Tables 4 and 5 show that most of the variables for the calf survival data and for some of the variables for culling data have a non-zero correlation indicating the proportional hazard assumption is violated for these variables. Moreover, the global test also showed a significant departure from zero.

Table 4. Score test of the proportional hazards assumption for the calf survival data

	rho	chisq	P
HYS	-0.1658	24.8521	0.001
AWTG	-0.05451	2.4377	0.013
CE	-0.08937	6.3677	0.012
WWG	0.12529	15.7168	0.001
ADG	-0.01047	0.1601	0.689
TPG	0.00747	0.043	0.835
NDI	-0.11659	11.6272	0.001
ADG2	0.05287	3.7149	0.053
AWTG2	0.04151	1.3899	0.238
Global		110.4742	0.000

HYS =herd-year-season; AWTG = arrival weight; WWG = weaning weight; ADG = average daily gain; TPG = total protein; NDI = number of disease incidence and CE = Calving ease score; ADG2 = average daily gain square; AWTG2 = arrival weight square

Table 5. Score test of the proportional hazards assumption for the culling data

	rho	chisq	P
HYS	-0.00875	22.1	0.638
HSV	-0.08924	21.9	0.001
SUP	0.01311	0.537	0.464
MILK	-0.27683	331.0	0.001
PROT	0.17949	84.4	0.001
FAT	0.15609	67.3	0.001
AFC	0.01691	0.872	0.352
FAT2	-0.3163	7.51	0.001
BCS	-0.00113	0.00443	0.947
Global		0.087	0.000

HYS = herd-year-season; HSV = herd size variation; SUP = herd supervision; MILK = milk production; PROT = protein production; FAT = fat production; FAT2 = fat production square; AFC = age at first calving; BCS = body condition score.

. The goodness of fit was assessed by plotting the Cox-Snell residuals against the cumulative hazards of Cox Snell residuals as presented in Figure 11. The graph shows that there is no evidence of systematic deviation from the straight line which indicate the adequacy of the model for the culling data set. However, the calf survival data showed a slight deviation from the straight line indicating the existence of some outlier observations.

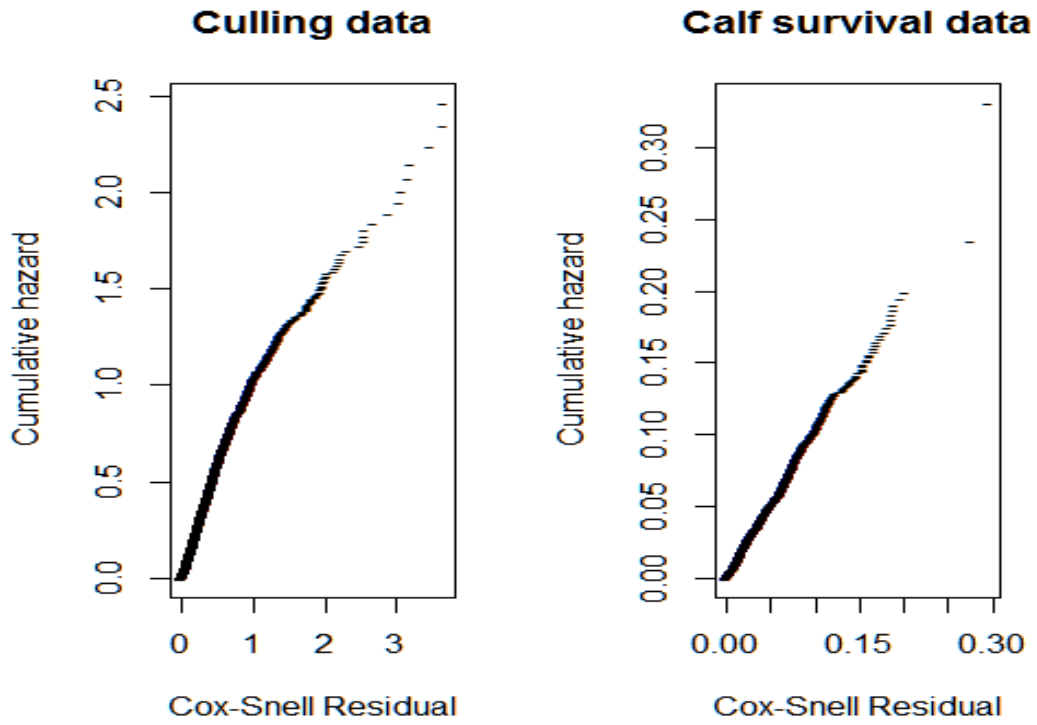


Figure 11. Cumulative hazards plot of the Cox Snell residual for the Cox-PH model

3.1.4. Accelerated Failure Time Model

One of the limitations of the Cox proportional hazards model is the assumption that the proportionality should remain constant regardless of the survival time of an individual. In this study the proportionality assumption did not hold for some of the variables for the two data sets (Tables 4 and 5). Therefore, the accelerated failure time model was used as an alternative model to the Cox model where the proportional hazard assumption is not held constant.

These data sets were analyzed using the different accelerated failure time such as Exponential, Weibull, Lognormal and Log-logistic models. The results are presented in Tables 6 and 7 for the culling data and calf survival data set, respectively.

The result of the AFT model can be interpreted as the size of the covariate effect in such way that in terms of difference in survival time among groups (Collett, 2004). For instance, in the present study the parameter estimate of the covariate, herd supervision from Weibull AFT model is -0.1744 (Table 6). This indicates that the survival time for the unsupervised herds is decreased by a factor 0.8399 compared to the supervised herds.

Table 6. Summary of results of fitting parametric AFT models to the culling data set.

Variable	Exponential		Weibull		Log-normal		Log-logistic	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	2.04223	0.1710	3.70029	0.111	1.699278	0.1840	-212.7853	21.785
HYS	0.00018	0.0000674	0.000149	0.00005	0.000093	0.000005	0.0163	0.0081
HSV	0.59303	0.0023	0.36700	0.0156	0.452833	0.0186	70.5197	3.5845
SUP	-0.2425	0.0231	-0.17801	0.0225	-0.204202	0.0313	-37.5363	6.1068
MILK	-0.00023	0.00002	-0.00017	0.000059	-0.000187	0.00008	-0.0388	3.9545
PROT	0.011909	0.000088	0.008540	0.00055	0.013089	0.00065	1.7393	5.4361
FAT	0.00216	0.00080	0.002020	0.00052	0.001358	0.00062	0.1940	5.2578
Fat2	-0.000025	0.0000026	-0.000019	0.0000014	-0.000019	0.0000016	-0.00236	0.00044
AFC	0.03378	0.0024	0.01910	0.00334	0.048757	0.01239	6.1547	3.7933
BCS	0.13784	0.012	0.100116	0.00708	0.123330	0.00913	18.2325	1.7859
log L	-21670.4		-21252		-21284.1		-21994	
AIC	43360.76		42525.95		42590.12		44010.01	

HYS = herd-year-season; HSV = herd size variation; SUP = herd supervision; MILK = milk production; PROT = protein production; FAT = fat production; FAT2 = fat production square; AFC = age at first calving; BCS = body condition score.

Table 7. Summary of results of fitting parametric AFT models to the calf survival data set.

Variable	Exponential		Weibull		Log-normal		Log-logistic	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	9.88228	9.882	10.16795	0.424059	10.45880	0.470196	2146.82	176.656
HYS	0.00404	0.000385	0.000288	0.000383	0.00319	0.000423	1.53	0.1605
CE	-0.11672	0.010338	- 0.11760	0.0108	-0.14328	0.011500	-0.1960	0.0235
AWTG	-0.01600	0.040215	-0.015554	0.004003	-0.01808	0.004478	-6.36	1.685
WWG	0.01597	0.002175	0.015511	0.002136	0.01713	0.002269	6.39	0.8784
TPG	-0.16722	0.073912	-0.012179	0.004587	-0.01328	0.005166	-5.27	1.9295
ADG	0.46945	0.107984	0.450720	0.020260	0.38793	0.021182	178.30	8.1838
ADG2	-0.106799	0.005911	-0.098288	0.0060	-0.13624	0.0074	-47.773	2.67
AWTG2	-0.000173	0.00021	-0.000156	0.00019	-0.00012	0.0002	-0.0759	0.0079
NDI	-0.24100	0.043407	-0.232457	0.041909	-0.28002	0.047434	-99.96	17.7725
log L	-7248.1		-7244.3		-7275.9		-7544.7	
AIC	14516.29		14510.54		15111.33		14573.83	

HYS =herd-year-season; AWTG = arrival weight; WWG = weaning weight; ADG = average daily gain; TPG = total protein; NDI = number of disease incidence and CE = Calving ease score; ADG2 = average daily gain square; AWTG2 = arrival weight square

Among the different AFT models it appears that the Weibull model had the lowest AIC for the two data sets (Tables 6 and 7) indicating that is the best fit among the different AFT models.

Assessment of influential observations i.e., the change in each regression coefficient when each observation is excluded from the data were calculated for each data set and some of the plots are presented in Figures 12 and 13. It appears that in both data set the there is no influential point as observed in both Figures. Additionally, observations with extreme value of residuals were excluded from analysis and parameter estimate were compared with the original estimate and the changes in parameter estimate were not significant.

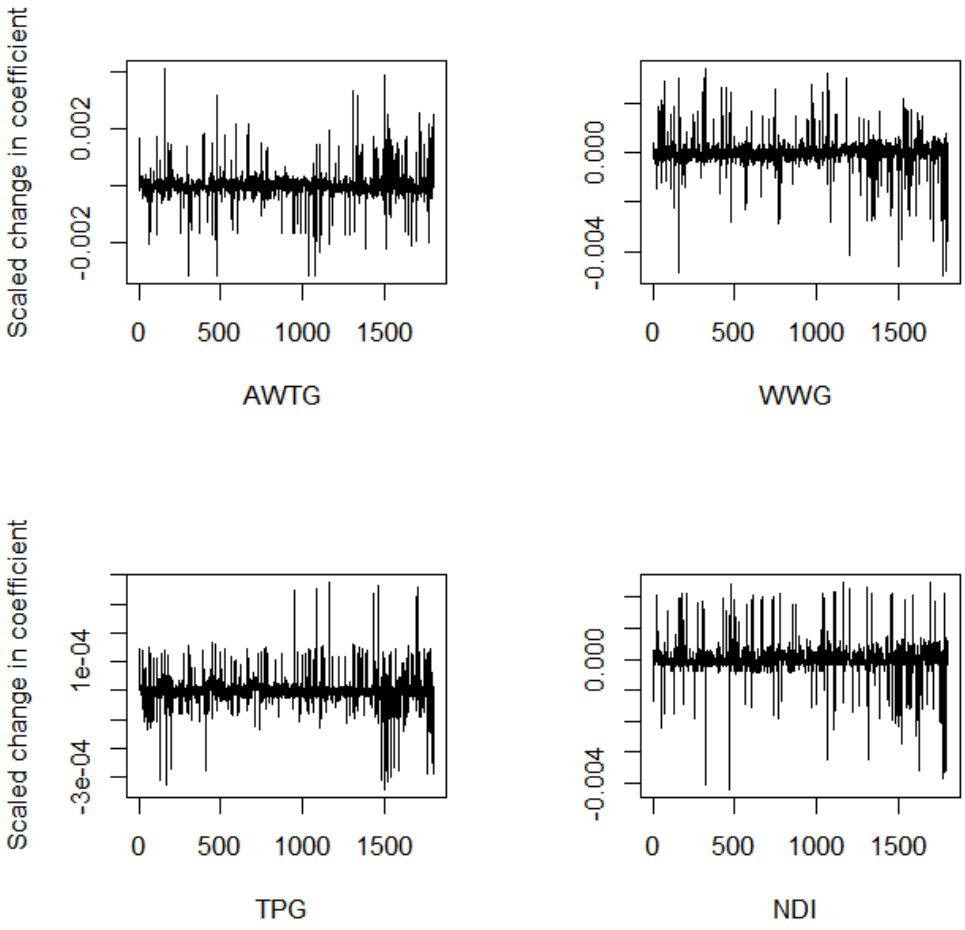


Figure 12. Change in each regression coefficient when each observation is removed from the data (influence statistics).

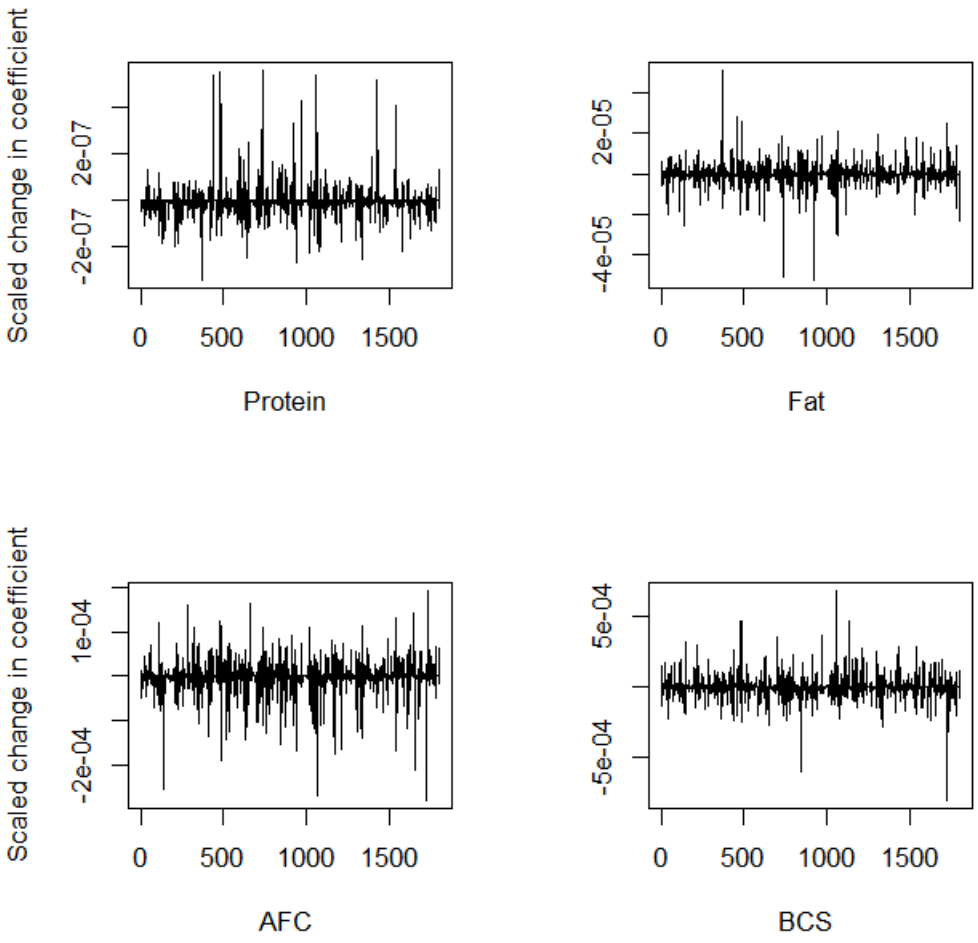


Figure 13. Change in each regression coefficient when each observation is removed from the data (influence statistics).

3.1.5. Partially Linear Single Index Survival Model

As discussed in the previous section, the partially linear single-index model has parametric and nonparametric components. Therefore, the application and implementation of the partially linear single-index survival model raises practical concerns of which covariates go into the nonparametric vector and which ones go into the parametric vector. In this study, first the subject matter knowledge related to each covariate and the underlying physiological mechanism that influences the ability of a calf to reach the next stage and the capacity of a cow to reach the second calving was utilized. Additionally, the subject matter knowledge was augmented with analyses of the data to investigate if the covariates were associated linearly or nonlinearly with the response variable.

For instance, for the culling data set as shown in Figure 6 one can infer that the relationship between age at first calving and body condition score (Figure 7) and the survival of cows is nonlinear showing that these variables have an intermediate optimum relationship with the response variable. Furthermore, the relationship between fat production and survival of cows is indicative of nonlinearity, where low and high producing cows tended to have higher risk of being culled. On the other hand, the average fat producing cows

tended to live longer than the other groups. Further, as an alternative way of examining the extent of nonlinearity was to define factor variable (Collett, 2004) to model the effect of fat production level on the hazard function. In this case a factor with ten levels of fat production was defined, where level 1 corresponds to the value of fat production lowest 10% of cows and level 10 corresponds to the value of fat production top 10% of the cow. The choice of levels corresponds to the quantiles of the distribution of the value of fat production (see R code Appendix B). This factor was fitted in the model by defining 9 indicator variables as fat2, fat3 and up to fat10. When the model containing the covariate fat and the rest of the variables the -2logL was 42,504 (Table 6). When the model containing fat2, fat3, up to fat10, the value of -2log L is 42,352 The change in the value of -2log L due to any nonlinearity is 152 on 8 df which is significant ($P< 0.001$). Therefore, it can be concluded that the effect of fat production on hazard of culling in this data set cannot be modelled using a linear function.

Figure 14 also plots the covariates (fat, age at first calving and body condition score) for the culling data, against the martingale residuals. The plots revealed that both covariates showed nonlinearity. In the absence of nonlinearity a flat line would have been observed (Therneau and Grambsch, 2000).

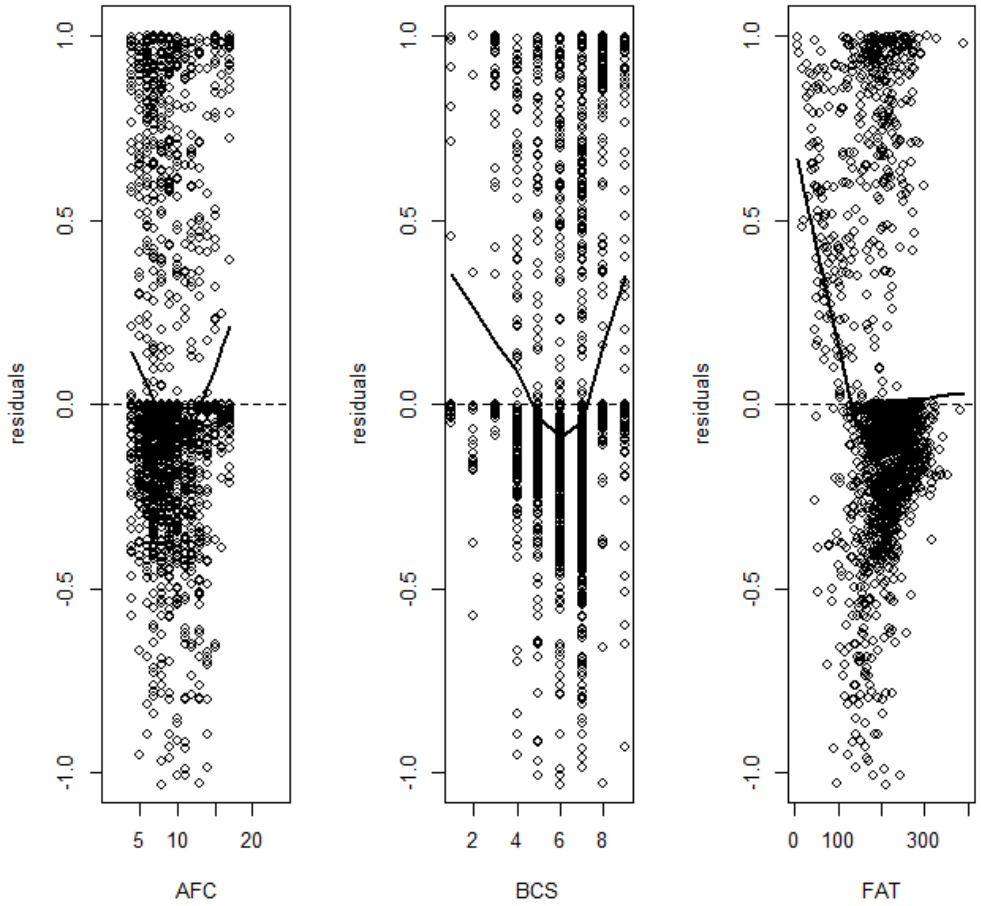


Figure 14. Martingale-residual plots for the covariates age at first calving (AFC), body condition score (BCS) and fat production for the culling data**.**

For the calf survival data, a similar pattern was observed for arrival weight (birth weight), total serum protein and average daily gain (Figure 15).

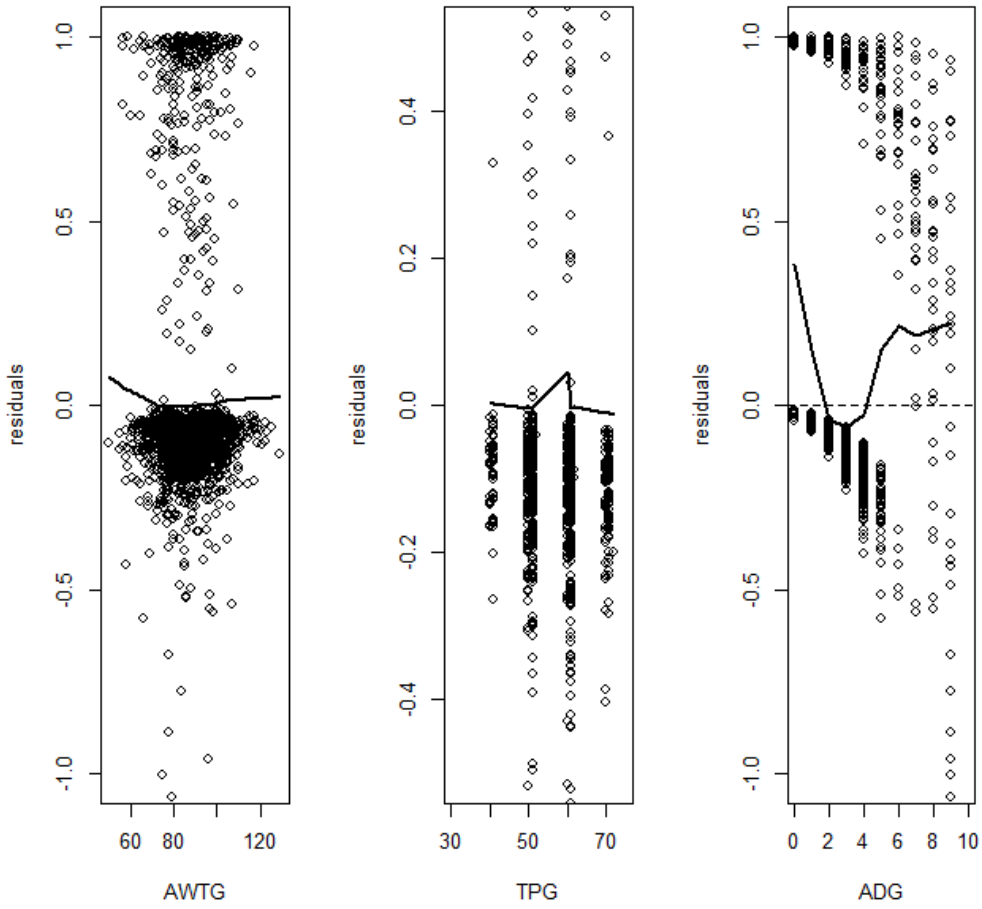


Figure 15 Martingale-residual plots for the covariates average daily gain for the calf survival data.

Therefore, for the culling data set the nonparametric component includes age at first calving, body condition score and fat production and the parametric component includes herd-year-season, supervision, milk production and protein production. For the calf survival data, arrival weight, total protein and average daily gain were included in the nonparametric component and herd-year-season, calving ease score, weaning weight group and number of disease incidence were in the parametric part of the model.

Estimate of parameters obtained from the partially linear single index survival model and Weibull model for the two data sets are presented in Tables 8 and 9. Table 8 shows that the estimates of the parameters in the parametric component (β) are similar under the ordinary Weibull linear model and the partially linear single-index survival model. However, the estimates of the nonparametric component (single-index) parameter, α, are different. This difference could be attributed largely to the nonlinearity of the estimated function. This nonlinearity relationship was also observed between survival time and those variables assigned to the nonparametric components of the two data sets.

Table 8. Estimates and standard errors of the parameters obtained from the Weibull model and the partially linear single index model for the culling data set

		Weibull		PLSISM	
		Estimate	SE	Estimate	SE
AFC	α_1	0.028018	0.0142	0.912131	0.23516
BCS	α_2	0.104516	0.0072	0.249126	0.10294
FAT	α_3	0.001553	0.00052	0.323143	0.14285
HYS	β_1	0.00013	0.00005	0.00031	0.00011
HSV	β_2	0.369476	0.0156	0.29672	0.13701
SUP	β_3	-0.174364	0.0225	-0.23145	0.01352
MILK	β_4	-0.000186	0.00005	-0.00278	0.00143
PROT	β_5	0.009755	0.00056	0.006256	0.00217

HYS = herd-year-season; HSV = herd size variation; SUP = type of milk recording; MILK = milk production; PROT = protein production; FAT = fat production; AFC = age at first calving; BCS = body condition score, PLSISM = partially linear single index survival model

Table 9. Estimates and standard errors of the parameters obtained from the Weibull model and the partially linear single index model for the calf survival data.

	Weibull		PLSISM	
	Estimate	SE	Estimate	SE
AWTG	-0.014354	0.003513	-0.34213	0.2101
TPG	-0.012076	0.00355	-0.55516	0.1783
ADG	0.51620	0.05720	0.76114	0.2254
HYS	0.000232	0.0000362	0.00031	0.0001
CE	-0.0722	0.0107	-0.10623	0.0435
WWG	0.013667	0.002031	0.015601	0.00139
NDI	-0.169772	0.040095	-0.27089	0.1534

HYS = herd-year-season; AWTG = arrival weight; WWG = weaning weight; ADG = average daily gain; TPG = total protein; NDI = number of disease incidence; CE = Calving ease score and PLSISM = partially linear single index survival model

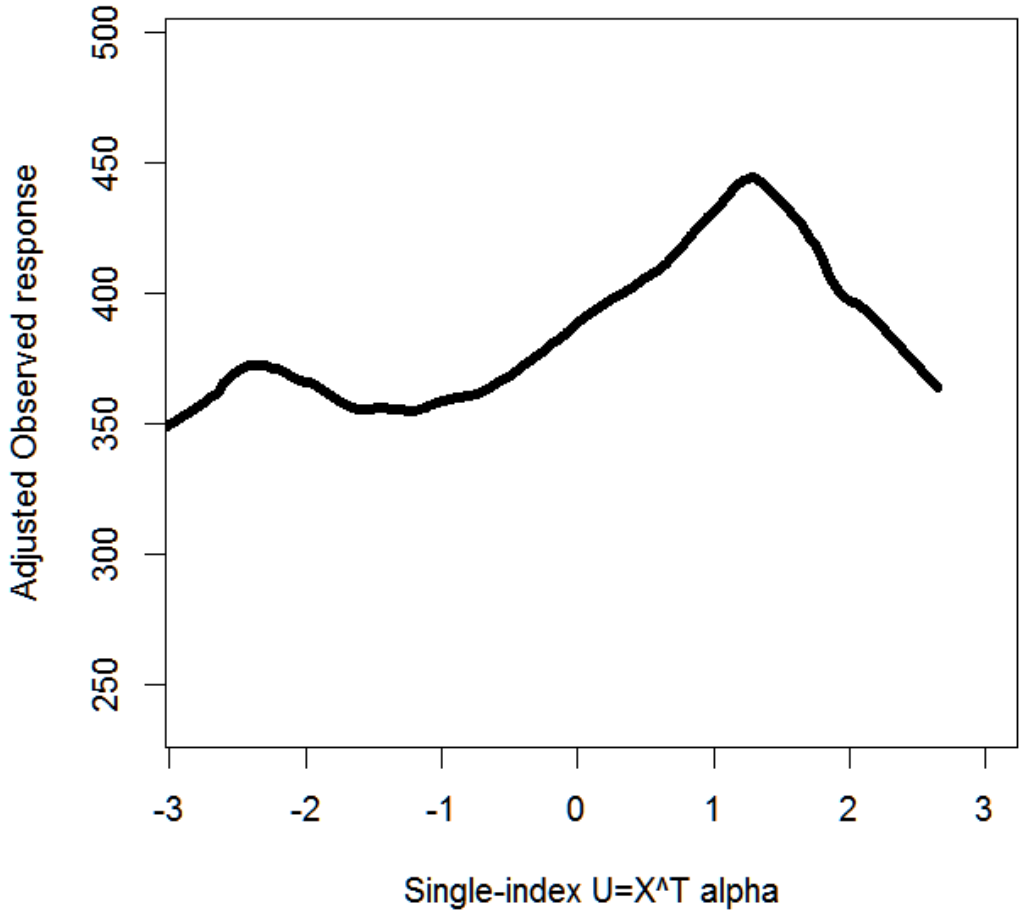


Figure 16. Observed response against the estimated single index value for the culling data

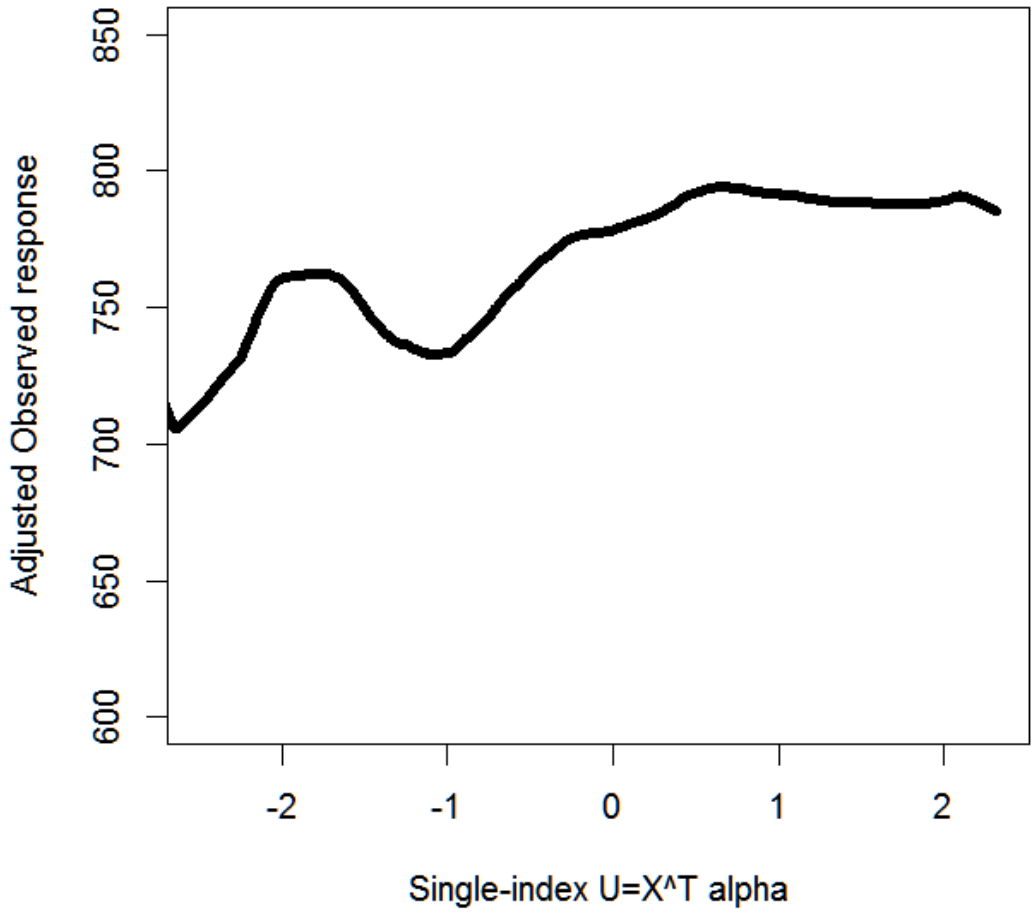


Figure 17. Observed response against the estimated single index value for the calf survival data

For the culling data, Figure 16 shows that at the beginning as the index increases, the survival of cows remain somewhat constant and then increases to the maximum. Once it reaches the maximum, as the index increases, the survival of cow decreases. In the previous section it was shown that as individual components of the index increase (age at first calving, body condition score and fat production), the survival of cows decreased. The possible biological and physiological explanation as to why age at first calving influences the survival of cows is that late calvings are presumably due to some problems associated with fertility or other health problems and these factors are likely to increase the risk of culling. Moreover, the economic consequence of delayed calving results in increased rearing costs hence, less profitability for the producers. On the other hand, cows that calve at younger age may encounter calving difficulty that in turn influences the subsequent health as well as productive capacity. With regard to body condition score, cows that are too thin (lower body condition scores) are more prone to metabolic problems and diseases, decreased production and poor fertility compared to cows that have adequate body weight and have higher body condition scores. However, fat cows have problem at calving since they are more prone to metabolic problems, such as dystocia or calving trouble, retained placenta, milk fever and ketosis. Age at first calving and body condition score is somehow related and this

relationship might affect the survival of cows. Therefore, this relationship may not be explained by the use of linear survival model.

A similar relationship but a different trend was observed for the calf survival data (Figure 17). As the index increases, the survival of calves increases in nonlinear fashion. At the beginning the effect is dramatic but once it reaches a threshold value it levels off. Looking at the individual components of the single-index such as the body weight, it reveals that calves with higher body weight tended to have higher risk of dying compared to the calves with lower body weight. Moreover, it was observed from the results that calves with high serum protein level have higher survival rate than those calves with lower volume. This is because the serum protein is important for immune response and ability to resist some disease incidence. The serum protein concentration in the blood is the reflection of how well the colostrum feeding and then absorption in to the blood stream was successful. Serum protein is also important to the calf for growth and development. These components of the index have somewhat intricate interdependence on each other that influences the calf survival in a complex way which may not be explained by application of the usual standard survival linear model.

Chapter 4

4. 1. Discussion and Conclusions

In the survival analysis literature, various statistical models have been proposed for analyzing censored data in the presence of covariates. Parametric and semiparametric survival models such as the accelerated failure time model and the Cox proportional hazards model are commonly used in many clinical trials, biomedical and agricultural studies. These models indicate the form of the conditional hazard function of survival time for a given set of covariates. An alternative approach is to use a direct relationship between survival time and its covariates by means of linear regression. For the last few decades the linear regression techniques for censored survival data have been used extensively because of the ease of interpretation of the results.

In the above-mentioned models, however, a relationship between survival time and the covariates is specified. Despite the fact that all these models have nice theoretical properties, they may not be flexible enough to describe the complexity of biological and physiological processes in many real data applications. The partially linear single index survival model as observed in the present study may have greater flexibility than other regression models in terms

of analyzing complex data since the link function in the model is assumed arbitrary.

In dairy cattle, the breeding goal is to increase lifetime profit per animal and per unit of time. Profit is a function of production and the time that a cow remains in herd (commonly called survival or longevity or herd life). Therefore, survival or longevity of cows is a trait of considerable economic importance since it has a significant impact on profitability. Increased longevity is associated with decreased culling and therefore decreased cost of raising or purchasing replacement females.

Several strategies have been suggested and used to analyze survival data in dairy cattle. These include a simple modeling of a 0-1 variable indicating whether the cow is still alive or dead at any specific time. In this approach, the response variable was considered as a binary trait and analyzed either using a linear or threshold model (VanRaden and Klaaskate, 1993; Jairath et al., 1998; Vollema and Groen, 1998; Boettcher et al., 1999; Sewalem et al 2007). Typically such type of data has a skewed distribution and analysis using traditional linear models may not be appropriate. Survival analysis using a proportional hazard model as suggested by Smith and Quaas (1984) is an alternative method for animal breeding survival data. Ducrocq et al. (1988) showed that proportional hazard models could be used for the analysis of length

of productive life. Ducrocq and Solkner (1998) developed the Survival Kit typically used by animal breeders for large populations using a Weibull model (Ducrocq, 2002; Sewalem et al., 2004) where several covariates are fitted as a linear effect.

Generally, in dairy cattle production there are several factors, covariates, that influence the survival of cows and those factors need to be accounted in the model in order to get reliable estimates. In this regard, Sewalem et al. (2005) studied longevity of Canadian dairy cows using censored linear regression model that included several covariates. In those analyses some of the continuous covariates were grouped and fitted to the model as a class effect. This grouping of covariates may also result in loss of information. In addition, some covariates may have nonlinear effects on the response variable. In this case, the traditional linear models fail to incorporate both linear and nonlinear covariate effects. The covariate effects in the current model are addressed in a semiparametric fashion, which offers better flexibility in modelling the relationship between the failure time and the covariates than the existing models. Hence, application of the current model is worthy of a full investigation using a larger data set that may include frailty model that accounts for the genetic effect of the animal.

The present study has demonstrated the application of partially linear single index model under random censorship using real data sets. The results have provided some insights which may be potentially useful in analysis of dairy cattle breeding data which describes the complex relationship between survival time and covariates. However, further studies should be carried out using large data set to validate the current results. Although this novel statistical approach is at its infant stage, a proper model checking procedure should be developed in the future

Future work

1. Consider an experimental trial consisting of a heterogeneous population of individuals which eventually, on treatment, divides into two groups. One group consists of those individuals who respond favourably to the treatment, appear subsequently free of any signs or symptoms of the disease, and may be considered immune or insusceptible to the disease and are said to be cured. The other group consists of those individuals who do not respond to the treatment and remain uncured. The widely used models in survival analysis are based on the assumption that every individual in the population under study will eventually experience the event of interest. However, this assumption cannot be used in some cases since some member of the population will never experience the event, regardless of how long they are observed. Therefore, it will be of great interest to apply the cure rate model using the current data to estimate the proportion of individuals who are permanently cured and analyze the effect of covariates on cure rate as well as the failure time of uncured individuals.

2. In dairy cattle breeding, it is crucial to identify those sires whose daughters are most resistant to a given disease or survived longer compared to other animals. In this field, several studies have been carried out using

proportional hazards models with the inclusion of a random genetic effect of sire. The proportional hazards model is powerful when the assumption of proportionality is true. However, this assumption is not always justifiable. Models based on first-hitting-time for Wiener process do not rely on those assumptions. Therefore, it will be of great interest and importance to apply this model using the existing data set.

3. The PLSISM has no likelihood and this makes it hard to compare this model with other models of interest. Therefore, in the future it would be of great interest to develop Akaike information criteria (AIC) or Bayesian information criteria (BIC) like quantities via the quasi-likelihood and apply some kind of penalty for effective number of parameters

References

Allaire, F. R. and Gibson, J. P. 1992. Genetic value of herd life adjusted for milk production. *J. Dairy Sci., 75:1349-1356.*

Boettcher, P. J, Jairath, L. K. and Dekkers, J. C. M. 1999. Comparison of methods for genetic evaluation of sires for survival of their daughters in the first three lactations. *J. Dairy Sci., 82:1034–1044.*

Bradburn, M. J., Clark, T. G., Love, S. B. and Altman, D. G. 2003. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *British J. Cancer, 89: 431–436.*

Box, G. E. P. 1987. Empirical Model Building and Response Surfaces, New York: John Wiley & Sons, Inc.

Buchinsky, M. and Hahn, J. 1998. An alternative estimator for the censored quantile regression model. *Econometrica, 66:653–671*.

Buckley, J. and James, I. R. 1979. Linear regression with censored data. *Biometrika, 66: 429–436.*

CDN: Canadian Dairy Network, 2011. (http://www.cdn.ca).

Carroll, R., Fan, J., Gijbels, I. and Wand, M. P. 1997. Generalized partially linear single-index models. *J. Amer. Statist. Assoc. 92:477–489.*

Clark, T. G., Bradburn, M. J., Love, S. B. and Altman. D. G. 2003. Survival analysis Part I: Basic concepts and first analyses. *British J. Cancer, 89:232–238.*

Chen, S., Dahl, G. B. and Khan, S. 2005. Nonparametric identification and estimation of a censored location-regression model. *J. Amer. Statist Assoc., 100:212–221.*

Collett, D. 2004. Modelling survival data in medical research. Chapman and Hall, London.

Cox, D. R. 1972. Regression models and life tables. *J. Royal Statist. Soc. B, 34: 187–220.*

Dabrowska, D. M. 1987. Nonparametric regression with censored survival time data. *Scandinavian J. Statistics, 14:181-192.*

Dekkers, J. C. M., Jairath, L. K. and Lawrence. B. H. 1994. Relationships between sire genetic evaluations for conformation and functional herd life of daughters. *J. Dairy Sci., 77:844–854.*

Ducrocq, V. and Solkner. J. 1998. “The Survival Kit V3.12”, a package for large analyses of survival data. *Pages 447–450 in Proc. 6th World Congress on Genetics Applied to Livestock Production, Armidale, Vol. 27.*

Ducrocq, V. Quaas, R. L., Pollak, E. J. and Casella. G. 1988. Length of productive life of dairy cows. 1. Justification of a Weibull model. *J. Dairy Sci., 71:3061–3070.*

Ducrocq, V. 2002. A piecewise Weibull mixed model for the analysis of length of Productive life of dairy cows. 7th World Congress on Genetics Applied to Livestock Production, August 19-23, 2002, Montpellier, France, *Session 20. Communication N° 20-04.*

Duncan, G.M. 1986. A semi-parametric censored regression estimator. *J. Econometrics, 32:5–24.*

Fan, J. and Gijbels, I. 1994. Censored regression: local linear approximations and their applications. *J. Amer. Statist. Assoc,, 89:560–570.*

Fan, J., Gijbels, I. and King, M. 1997. Local likelihood and local partial likelihood in hazard regression. *The Annals of Statist, 25:1661-1690.*

Fernandez, L. 1986. Non-parametric maximum likelihood estimation of censored regression models. *J. Econometrics, 32: 35–57.*

Henderson, L. 2009. A genetic analysis of dairy cattle health traits and survival. MSc thesis, 2009. University of Guelph.

Heuchenne, C. and Van Keilegom, I. 2007. Location estimation in nonparametric regression with censored data. *J. Multivariate. Analysis, 98: 1558-1582.*

Härdle, W. and Stoker, T.M. 1989. Investigating smooth multiple regression by the method of average derivative. *J. Amer Statist Assoc., 84:986–995.*

Horowitz, J. L. 1988. Semiparametric M-estimation of censored linear regression models. *Advanced Econometrics, 7:45–83.*

Ichimura, H. 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics, 58:71–129.*

Jairath, L., Dekkers, J. C. M., Schaeffer, L., Liu, Z., Burnside, E. B. and Kolstad, B. 1998. Genetic evaluation for herd life in Canada. *J. Dairy Sci., 81:550-562.*

Jarrow, R. A. and Turnbull. S. M. 2000. The intersection of market and credit risk. *J. Banking Finance, 24:271–299.*

Kalibfeisch, J. D. and Prentice, R. L. 1980. The Statistical analysis of failure time data. John Wiley & Sons , New York.

Kaplan E. L, Meier P. 1958. Nonparametric estimation from incomplete observations. *J Amer. Statist. Assoc., 53: 457–81.*

Kleinbaum, D. G. and Klein, M. 2005. Survival analysis, Statistics in health sciences, Springer-Verlag, New York.

Koul, H., Susarla, V. and Van Ryzin, J. 1981. Regression analysis with randomly right-censored data. *Ann. Statist., 9:1276–1288.*

Lai, T. L., Ying, Z. and Zheng, Z. 1995. Asymptotic normality of a class of adaptive statistics with applications to synthetic data methods for censored regression. *J. Multivariate Analysis, 52:259–279.*

Lawless, J.F. 2003. Statistical models and methods for lifetime data, Wiley, New York.

Lewbel, A. 1998. Semiparametric latent variable model estimation with endogenous or mismeasured regressors. *Econometrica, 66:105–121.*

Lewbel, A. and Linton, O. 2002. Nonparametric censored and truncated regression. *Econometrica, 70:765–779.*

Liang, H. and Zhou, Y. 1998. Asymptotic normality in a semiparametric partial linear model with right-censored data. *Comm. Statist. Theory & Methods, 27:2895–2907.*

Lu, X. and Cheng, T. L. 2007. Randomly censored partially linear single-index models. *J. Multivariate Analysis, 98:1985-1922.*

Lu, X. and Burke, M. D. 2005. Censored multiple regression by the method of average derivatives. *J. Multivariate Analysis, 95:182–205.*

Lu, X., Chen, G., Song, X. K. and Singh, R.S. 2006. A class of partially linear single-index survival models. *Canadian J. Statist., 34:99–116.*

Nardi, A. and Schemper, M. 2003. Comparing Cox and parametric models in clinical studies. *Statist. Med., 22:3597–3610.*

Newey, W. K. and Stoker, T.M. 1993. Efficiency of weighted average derivative estimators and index models. *Econometrica, 61:1199–1223.*

Nielsen, J.P. and Linton, 0. B. 1995. Kernel estimation in a nonparametric marker dependent hazard model. *The Annals of Statistics, 23:1735-1748.*

Powell, J. L.1986. Censored regression quantiles. *J. Econometrics, 32:143–155.*

Sewalem, A., Kistemaker, G. J., Miglior, F. and Van Doormaal, B. J. 2004. Analysis of the relationship between type traits, inbreeding and - functional survival in Canadian Holstein Dairy Cattle. *J. Dairy Sci., 87: 3938-3946.*

Sewalem, A., Miglior, F. Kistemaker, G. J., Sullivan, P. Huapaya G. and Van Doormaall. B. J. 2007. Modification of Genetic Evaluation of Herd life from a 3-trait to 5-trait Model in Canadian dairy cattle. *J. Dairy Sci., 90:2025-2028.*

Sewalem, A., Kistemaker, G. J. Ducrocq, V. and Van Doormaal. B. J. 2005. Genetic analysis of herd life in Canadian dairy cattle on a lactation basis using a Weibull proportional hazard model. *J. Dairy Sci., 88: 368-375.*

Shoukri, M. M, Attanasio, M. and Sargeant. J. M. 1998. Parametric versus semi-parametric models for the analysis of correlated survival data: A case study in veterinary epidemiology. *J. Applied Statist., 25:357-374.*

Singh, R.S. and Lu, X. 2002. Censored additive regression models, Handbook of applied econometrics and statistical inference, in: A. Ullah, A.T.K. Wan, A. Chaturvedi (Eds.), Statistics: *Textbooks and Monographs, vol. 165, Dekker, New York, 143–157.*

Smith, S. P. and Quaas. R. L 1984. Productive lifespan of bull progeny groups: failure time analysis. *J. Dairy Sci., 67:2999–3007.*

Therneau, T.M. and Grambsch, P.M. 2000. Modeling of survival data: extending the Cox model. New York: Springer-Verlag.

VanRaden, P. M. and Klaaskate. E. J. H. 1993. Genetic evaluation of length of productive life including predicted longevity of live cows. *J. Dairy Sci., 76:2758–2764.*

Vollema, A. R. and Groen. A. F. 1998. A comparison of breeding value predictors for longevity using a linear model and survival analysis. *J. Dairy Sci., 81:3315–3320.*

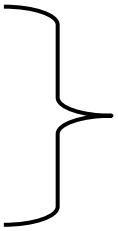
Wang, Q. H. and Zheng, Z. G. 1997. Asymptotic properties for the semiparametric regression model with randomly censored data, *Sci. China Ser. A 40:945–957.*

Appendix.

A

```
#R-code that Calculate the SE of parameters using Bootstrap approach
rm(list=ls())
options(width=60,length=200)
library(survival)
library(pspline)
library(MASS)

newdata11  <- read.table(paste('G:/path....', "data_file.txt", sep="/"), header=T)
  head(newdata11)

# Initialize parameters

alp1 <- NULL
alp2 <- NULL
alp3 <- NULL
bet1  <- NULL
bet2  <- NULL
bet3  <- NULL
bet4  <- NULL
bet5  <- NULL
reps=500

for (Nout in 1:reps)      # loop over the entire program
```

```
{
 newdata1<-newdata11[sample(1:nrow(newdata11),replace=T),]
    .
    .
    .
    .
```



PLSISM

```
### Save the final output of each round of bootstrap
              bet1[Nout] <- bEst[1,1]
              bet2[Nout] <- bEst[1,2]
              bet3[Nout] <- bEst[1,3]
              bet4[Nout] <- bEst[1,4]
              bet5[Nout] <- bEst[1,5]
             alp1[Nout] <- aEst[1,1]
              alp2[Nout] <- aEst[1,2]
              alp3[Nout] <- aEst[1,3]
   #   cat("This is Bootstrap #", Nout, "bet1=", bet1,";", "alp1=", alp1, "\n")
}


summary(bet1)
summary(bet2)
summary(bet3)
summary(bet4)
summary(bet5)
```

```
sebt1= sqrt(var(bet1))

sebt2= sqrt(var(bet2))

sebt3= sqrt(var(bet3))

sebt4= sqrt(var(bet4))

sebt5= sqrt(var(bet5))

sealp1= sqrt(var(alp1))

sealp2= sqrt(var(alp2))

sealp3= sqrt(var(alp3))

##########################################
```

B

R-Code that determines the factor level for fat production based on Collett, 2004

```
rm(list=ls())

options(width=60,length=200)

library(survival)

library(pspline)

library(MASS)


newdata1   <- newdata11  <- read.table(paste('G:/path....', "data_file.txt", sep="/"),

head(newdata1)

attach(newdata1)


dat <- data.frame(fat, group=findInterval(fat,quantile(fat, prob=c(0, 0.1,0.2,0.3,0.4,
0.5,.6,0.7,0.8,0.9 ,1)), all.inside=TRUE))

dat$gmean <- ave(dat$fat, dat$group)

dat
```