Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 69(2) 2015 149-160

Analysis of Calf Survival and Culling Data Using a Partially Linear Single Index Model

A. Sewalem[1, 2], A.F. Desmond[3], R.S. Singh[3] and X. Lu[4]
[1]*Guelph Food Research Centre, Agriculture and Agri-Food Canada, Guelph, ON, N1G 5C9, Canada*
[2]*Canadian Dairy Network, Guelph, ON, Canada, N1K 1E5*
[3]*Department of Mathematics and Statistics, University of Guelph, Guelph, ON, N1G 2W1, Canada*
[4]*Department of Mathematics and Statistics, University of Calgary, Calgary, AB. T2N 1N4, Canada*

Received 25 September 2014; Revised 16 April 2015; Accepted 27 April 2015

SUMMARY

In many practical situations the linear model is not complex enough to capture the underlying relationship between the response variable and its regressors. This paper explores this association in dairy cattle data using the partially linear single-index survival model (PLSISM). In addition, parametric accelerated failure time (AFT) survival models were also used. Calf survival and culling data (survival from first calving to second calving) sets were used. The calf survival data contains, as covariates, arrival weight, weaning weight, total serum protein, calving ease score, herd-year-season of calving and number of disease incidences. The culling data set includes: age at first calving, body condition score, level of production (milk, fat and protein yield), herd size variation, type of milk recording and herd-year-season of calving. For the calf survival data, arrival body weight, weaning weight, total serum protein and age at first calving were included in the nonparametric component of the PLSISM. For the culling data, body condition score and fat production were included in the nonparametric component of the PLSISM. All factors included in the respective models of the two data sets had a statistically significant effect on the survival time. The simpler AFT model provides an intuitively more simple interpretation of covariate effects. Indeed, estimates of the parametric component were similar for the two models of the two data sets. However, the estimates of the nonparametric component differed from the parametric analysis. This difference may be attributed largely to the nonlinearity of the estimated function, suggesting that the standard linear model did not adequately capture the underlying association between the response and regressors in this study.

Keywords: Survival, Nonlinear, Accelerated failure time model, Dairy cattle.

1. INTRODUCTION

In studying the relationship between a response and a set of predictor variables, the mean response variable is often assumed to be a linear regression function of the regressors. In many practical situations, however, the linear model is not complex enough to capture the underlying relationship between the response variable and its associated covariates. Indeed, 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. For the past decades, several models have been developed to study high dimensional data by 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.

Corresponding author: Asheber Sewalem
E-mail address: sewashm@yahoo.com

In the references cited, 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 (1972) proportional hazards model. The Cox (1972) regression model plays a central role in survival analysis in which the conditional hazard of failure at time t given the covariate vector Z takes a semi-parametric form

$$h(t : Z) = h_0(t) \exp(\beta^T Z),$$

where the baseline hazard $h_0(t)$ is an unspecified function of time t, and the covariate effects are specified in such a way that the parameter vector β represents the log-linear effects on the hazard function. 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; examples include, Dabrowska (1987) and Nielsen and Linton (1995), where the hazard function $h(t, z)$ is completely unspecified. Fan *et al*. (1997) relaxed the fully nonparametric specification to the form $h(t, z) = h_0(t)$ $\lambda(z)$ where both h_0 and λ are nonparametric functions.

Fan and Gijbels (1994) proposed a censored nonparametric regression estimator based on a class of unbiased data transformations using only univariate regressors. Wang and Zheng (1997) and Liang and Zhou (1998), however, extended this to multiple regressors. Further, Singh and Lu (2002) studied censored nonparametric additive regression models based on some special data transformations. On studying the estimation of an unknown multiple regression function, 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 survival analysis, an alternative model to the proportional hazards model or the multiplicative hazards model is the accelerated failure time model (Lawless 2003).

The 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 proportional hazards survival model has been extensively used in dairy science for several years (Ducrocq 2002, Sewalem *et al.* 2005). The accelerated failure time model assumes the logarithm of the failure time is a linear function of covariates included in the model plus a random error term (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 *iis* is non-Gaussian, (2) the response is subject to censoring and (3) the dependent variable, Y or log T, is the logarithm of time.

When there is no censoring, the ordinary least squares method can be used to obtain point estimates of parameters, where the dependent variable is log T. However, the ubiquity of censoring in survival data makes it difficult to use the above procedure. Nevertheless, the maximum likelihood procedure with different distributional assumptions on the random error is still applicable for estimation purposes (Lawless 2003).

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 is assumed that the effect of a covariate is to multiply the hazard by some (possibly covariate-dependent) constant, whereas, in the latter case, it is assumed that the effect of a covariate is to multiply the predicted event time by some constant (Kalbfleisch and Prentice 1980, Lawless 2003, Collett 2004).

Lu and Cheng (2007) investigated a class of partially linear single-index models under random censoring as a class of accelerated failure-time models without the specification of the distribution function of the response variable. Sewalem (2012) expanded the above-mentioned model to handle several covariates and added a bootstrap procedure to calculate the standard error of parameters. To date, in the case of survival analysis in dairy cattle, the covariates are often fitted in the model as linear multiplicative effects,

although several covariates are believed to have nonlinear effects on the survival of a cow. Therefore, the objectives of this study were: a) to use the accelerated failure time model to analyze survival data instead of the proportional hazards model; and b) to explore the association of the response variable and various regressors in dairy cattle breeding data using the partially linear single-index model extension of the accelerated failure time model.

2. MATERIALS AND METHODS

2.1 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 below. Calf survival data was obtained from a commercial rearing facility located in New York, USA (Henderson *et al.* 2011). The centre raises heifers to varying target ages based on the contractual arrangement between the facility and the farm of origin or producers (Henderson *et al.* 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, the following information was recorded: 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.

Some of the fixed continuous variables were grouped into intervals and fitted as a factor effect in the model. This appears to be a common practice in dairy science, although an interesting question is to what extent information is lost. 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 = disease occurrence, 2 = 2 disease occurrences, 3+ = 3 or greater disease occurrences) farm of origin or herd and sire.

The culling data was obtained from lactation and type classification records extracted for the Canadian May 2011 genetic evaluation of the Holstein breed (CDN 2011). Length of productive lifetime 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 complete (uncensored) if the cow received a termination code, indicating that the cow was removed for any reason from the herd. Records associated with missing identification, incorrect calving dates and age at first calving (AFC) outside 18-40 months range were excluded from the analysis. Type information consisted of phenotypic type scores such as body condition score, a descriptive trait evaluated on an ordinal scale from 1 to 9. Body Condition Scoring (BCS), recorded during the first lactation period, 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 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 (year of calving was from 2005 to 2010; 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 less than 5%; nearly unchanged, if the change is within a 5% decrease to a 10% increase; and increasing for an increase in herd size of more than 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 for cows producing less than 0.3 standard deviations below the herd-year average, average for cows producing between 0.3 standard deviations below and 0.5 standard deviations above the herd-year average and high for cows

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

The primary goal of the study was 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. The data were analyzed using the following models

2.2 Parametric Model

The AFT model we consider is given by:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \sigma\varepsilon_i,$$

where Y_i is the log survival time, β_0 is the intercept, σ is the scale parameter and ε_i a random error term has a standard extreme value distribution and a set of k covariates (x_1, x_2, ..., x_k), whose impact is measured by the size of the respective coefficients (β_1, β_2, ..., β_k). For the culling data the covariates included were, age at first calving, body condition score, herd size variation, type of milk recording (supervised versus unsupervised herds), level of production (milk, fat and protein) and sire. After editing, the numbers of records were 8804 from 452 herds sired by 1285 bulls. For the calf survival data, the covariates included were arrival weight, total serum protein, average daily gain, calving ease score, herd-year-season, number of disease incidences and sire. After editing, the data set consisted

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

	Culling data		Calf survival data	
	Censored	Failure	Censored	Failure
Number of records	5,616	3,188	6,694	678
Average time (days)	318	208	802	365
Minimum time (days)	6	1	800	46
Maximum time (days)	600	425	807	768

of 7,372 Holstein calf records from 36 herds sired by 264 bulls.

2.3 Semiparametric Model

In addition to the above-mentioned model, the following partially linear single-index survival model (PLSISM) was used to analyze the data (Lu and Cheng 2007),

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

where Y is the log survival time, V and X are the associated covariate vectors of dimensions q and p respectively. For ease of understanding the covariates are 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-dimensional 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-dimensional vector parameter α_0. $\sigma(\cdot,\cdot)$ is the conditional variance representing possible heteroscedacity; $\|\cdot\|$ denotes the Euclidean norm. The constraint $\|\alpha_0\| = 1$ on the single-index coefficient parameters is required for parameter identifiability. We assume that (V, X) and ε are independent, and that $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = 1$. Let C be the random censoring time associated with the log survival time Y. We 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).

Application of the partially linear single-index approach using the above mentioned model raises practical issues of which covariates go into the nonparametric vector V and which ones go into the parametric vector X. In this study, we utilized the subject matter knowledge related to the individual calf and the underlying physiological mechanism that influences the ability of a calf to reach the next stage. In addition, preliminary analysis was carried out to investigate if the covariates were associated linearly or nonlinearly with the response variable. For the culling data set, the nonparametric component includes BCS, AFC and level of fat production. For the calf survival data, arrival weight, weaning weight and total serum protein are included. The other covariates are included in the parametric component.

2.3 Estimation Procedure

Since the distribution of the error term in the model is not specified, deriving the full likelihood function for 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 assumptions are made about the distribution of the response Y.

Let $\theta = (\alpha, \beta)$ be the vector of model parameters and if the data is fully observed, *i.e*., $Z \equiv Y$ the quasi likelihood estimator of θ_0, = (α_0, β_0) and λ_0 are the minimizers of the following quasi-likelihood function of $\{(V_i, X_i, Z_i, \delta_i)$: $i = 1, ..., n\}$,

$$\ell_n(\theta, \lambda) = \sum_{i=1}^{n} [\boldsymbol{Y}_i - \{\beta^T \boldsymbol{V}_i + \lambda(\alpha^T \boldsymbol{X}_i)\}] \text{ with } \|\alpha\|=1$$

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_0. To overcome these difficulties, first synthetic data or pseudo responses were produced using:

$Z_{i\hat{G}} = (1+\phi)L_{i\hat{G}} - \phi K_{i\hat{G}}$ (following the procedure of Lu and Cheng 2007).

Here $L_{i\hat{G}} = \int_{-\infty}^{\infty} (\frac{I[Z_i \geq s]}{(1-\hat{G}(s-))} - I[S<0])ds$,

$\boldsymbol{K}_{i\hat{G}} = \frac{Z_i \delta_i}{(1-\hat{G}(Z_i-))}$, ϕ is a tuning parameter which control the weights put on censored and uncensored observations and $I(.)$ is the indicator function. $(1-\hat{G}(.-))$ is the left continuous version of Kaplan-Meier estimator defined by

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

$Z_{(1)} \leq Z_{(2)} ... Z_{(n)}$ are the order statistics of Z-sample and δ_i is the associated δ with $Z_{(i)}$, i = 1, 2, ..., n. The observed data $(\boldsymbol{V}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \delta_i)$ is replaced by $(\boldsymbol{V}_i, \boldsymbol{X}_i, \boldsymbol{Z}_{i\hat{G}})$. 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, we may substitute the Kaplan-Meier estimator $\hat{G}$ for G. The transformation is still asymptotically unbiased in this case. This class of transformations was introduced by Fan and Gijbels (1994), and Koul *et al.* (1981). Using the transformed data, both the parametric component (β_0) and the nonparametric component (λ_0) were estimated by iteratively applying a local linear fit to the quasi log likelihood.

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 $\lambda_0(\cdot)$ and the parametric components $\theta_0 = (\alpha_0, \beta_0)$. The iterative algorithm consists of the following steps:

Step 1. Treat the pseudo-responses $\boldsymbol{Z}_{i\hat{G}}$ as complete data and apply the estimation procedure for the partially linear single-index models, to obtain initial estimates $\hat{\alpha}$ and $\hat{\beta}$ of α_0 and β_0 respectively, with the restriction $\hat{\alpha} = 1$ 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 as described by Lu and Cheng (2007).

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

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

Step 4. Cycle Steps 2 and 3 until convergence of $\hat{\theta}$.

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 model was fitted and the corresponding parameters were estimated along with the nonparametric component as described in the above estimation procedure.

3. RESULTS AND DISCUSSION

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 are presented in Table 1. The percent censored and uncensored records for the culling data

set were 63.8 and 36.2%, respectively. The corresponding figures for the calf survival data were 90.8 and 9.2%. The censoring proportion for the calf survival data was higher than the culling data set.

Analysis of the culling data shows that all effects included in the model have a statistically significant effect ($P < 0.01$). The preliminary analysis also shows that there is no significant interaction among the main effects. The quadratic effect of level of fat production was significant ($P < 0.01$) and, therefore, it was included in the final model.

A detailed analysis of the effect of each predictor variable on the survival of calves 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 an accelerated factor, defined as the ratio between estimated survival time under the influence of certain predictors of survival and the average survival time in a reference group, which is set to one as a bench mark. An accelerated factor greater than one indicates a beneficial effect of a predictor variable on survival of cows. A value less than one indicates a detrimental effect for a given predictor on survival. Hence, for the culling data, the annual change in herd size was associated with relatively longer survival time in expanding herds compared to stable herds. The survival time for expanding herds increased by a factor of 1.29 and 1.59 compared to those herds which are stable or shrinking, respectively. For instance, if the median survival time of the stable herds were 400 days, then the estimated median survival time of expanding and shrinking herds would be 516 and 324 days, respectively.

The survival time associated with the type of milk recording supervision shows that cows in unsupervised herds had survival times lower by a factor of 0.84 compared to supervised herds. The effect of within herd-year production deviations (milk, fat and protein) had a significant influence on the survival time of cows. The estimated accelerated factor for cows producing 0.3 standard deviations below the herd-year mean had lower survival time, by a factor of .75, than the average producers for milk, while high producing cows had longer survival times by a factor of 1.19 compared to the average producers. The influence of within herd-year protein yield deviations follows the same trend as that of milk yield (0.63 and 1.34 for low and high producing cows compared to the average producing cows). The accelerated factors for cows producing 0.3 standard deviations below the herd-year-mean had lower survival time than average producers for protein. With regard to fat production, cows producing 0.3 standard deviations below the herd mean were 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 lower survival times compared to the average producing cows.

The effect of age at first calving also had a significant influence on the survival of cows to second calving (Fig. 1). The accelerating factor was lower for older heifers and young 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 decrease the survival time of cows. Moreover, cows with delayed calvings are less profitable due to higher rearing costs. Fig. 1 also shows a trend towards a higher risk of culling for cows with first calving at less than 21 months of age, as younger cows may be at a greater risk of having calving difficulties or dystocia.

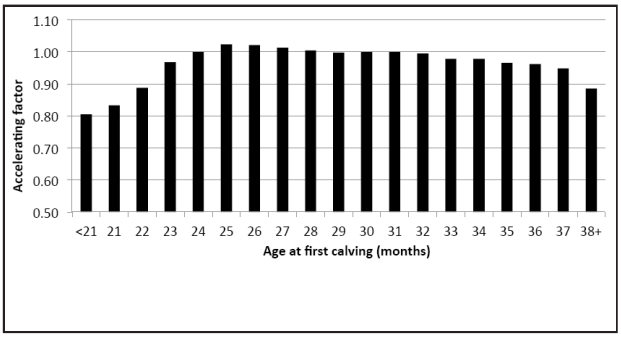

Fig. 1. Estimates of the accelerating factors for age at first calving (AFC at 24 was set =1)

Fig. 2 shows a clear nonlinear relationship between body condition score and survival of cows to the second calving. Survival times of cows with a score of 1 (classified as lean) were smaller by a factor of 0.58 compared to the reference class (score of 5). Similarly, cows with a score of 9 (classified as fat) were less likely to stay longer in the herd compared to the reference group. Overall, Fig. 2 does show that body condition score is an intermediate optimum trait indicating 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 (Fig. 1) and body condition score (Fig. 2) with the survival of cows

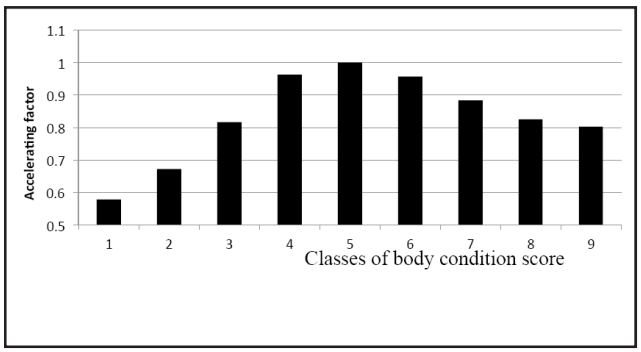

Fig. 2. Estimates of the accelerating factor for body condition score (BCS of 5 was set to 1)

to the second calving is somewhat nonlinear and the application of parametric linear survival regression 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. Further, as an alternative way of examining the extent of nonlinearity we used a factor variable along the lines of (Collett 2004) to model the effect of fat production level on the survival function. In this case a factor with ten levels of fat production was defined, where level 1 corresponds to the lowest decile of fat production and level 10 corresponds to the top decile of fat production. Thus, the choice of levels corresponds to appropriate deciles of the distribution of fat production. This factor was fitted in the model by defining 9 indicator variables as fat2, fat3, up to fat10. When the model containing the covariate fat and the rest of the variables was fit the deviance -2logL was 42,504. On the other hand, when the model containing the factor variable for fat is used, the value of -2log L was 42,352. The change in the value of -2log L due to any nonlinearity is 152 on 8 df which is highly significant ($P< 0.001$). Therefore, we conclude that the effect of fat production on survival of cows has a nonlinear effect in this data set and should not be modeled using a linear function.

A closer look at the effect of each predictor variable on calf survival indicated significant differences within each categorical variable on calf survival. For instance, for the calving ease covariate, the accelerated factors for calves born with easy pull and hard pull or surgery, were 0.91 and 0.83, respectively, compared to unassisted calving (Fig. 3). The influence of calving ease score on survival of calves might imply that as calving difficulty increases,

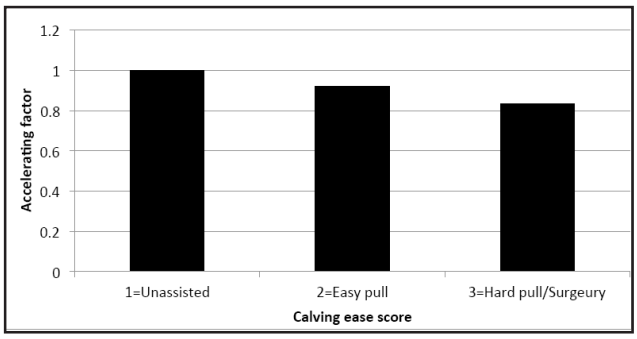

Fig. 3. Estimated accelerating factor of calving ease score

calves experience distress and 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 the survival time of calves. As total protein group score increased, the survival time of calves increased. The accelerating factor for total protein groups 1, 2, 3 and 4, were 1.00, 1.24, 1.3 and 1.62, respectively.

Weaning weight of calves influenced survival of calves significantly ($P < 0.001$). Fig. 4 exhibits the survival rate for weaning weight classes. The estimated accelerating factor of heifers weighing between 75-111 lbs at weaning was 0.68 times lower than a heifer weaned at an average weight of 131-150 lbs. On the other hand, heifers weaned in the heaviest weaning weight class (>175 lbs) were approximately 120% more

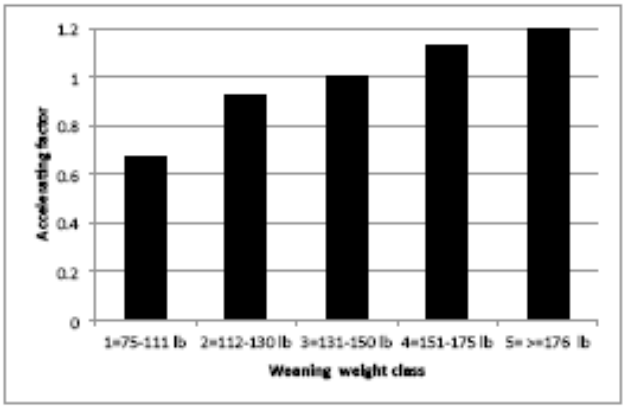

Fig. 4. Estimated accelerating factor of weaning weight class

likely to survive than heifers weaned in the average weaning weight class. Heifers with heavier weaning weights were more likely to urvive 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.

Arrival weight also influenced the survival of calves during the study period. Calves in the low arrival weight group were less likely to survive compared to the average arrival weight group (Fig. 5). Similarly, the higher arrival weight group was also found to have lower survival times compared to the average group. The shorter survival time associated with higher birth weight might be associated with calving difficulty resulting in more health issues in early life, thus increasing the risk of mortality. With regard to the arrival weight (birth weight), as presented in Fig. 5 there appears to be an optimal weight class (84-92 lbs) in relation to the survival of calves.

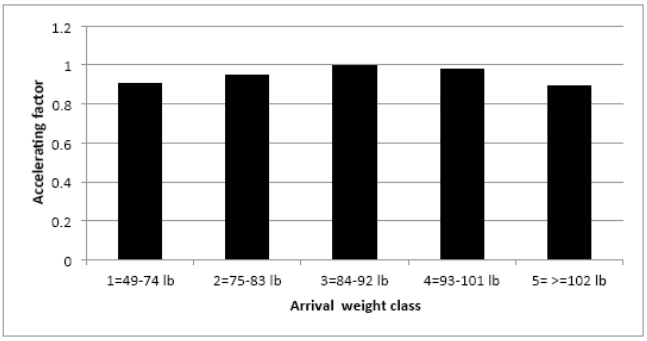

Fig. 5. Estimated accelerating factors of arrival weight class

Average daily gain also influenced the survival of calves till exit. For instance calves that had low average daily gain were at 44% 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 that had two or three disease incidences had a considerably increased relative risk of short survival time compared to calves that had no disease at all. For instance, the accelerated factor for calves experiencing 2 or 3 or more disease occurrences requiring treatment prior to weaning were 0.74 and 0.48 relative to calves that experienced no preweaning disease (Fig. 6).

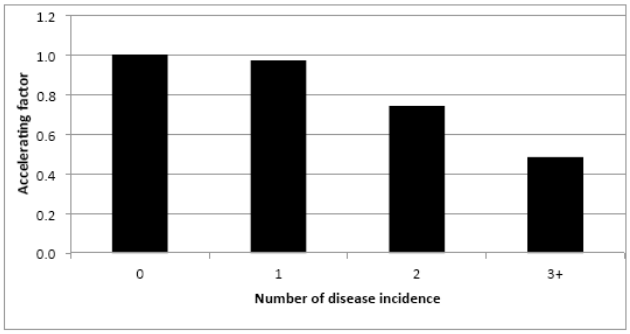

Fig. 6. Estimated accelerating factor for number of disease incidences.

3.1 Semiparametric 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 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, 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 Fig. 1, one can infer that the relationship between age at first calving and body condition score (Fig. 2) and the survival of 2 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. Additionally, plots of covariates (fat, age at first calving and body condition score) for the culling data, against martingale residuals (not shown) revealed that both covariates showed nonlinearity. For the calf survival data, a similar pattern was observed for arrival weight (birth weight), total serum protein and average daily gain.

Therefore, for the culling data set the nonparametric component included age at first calving, body condition score and fat production and for the calf survival data, arrival weight, total protein and average daily gain. The rest of the predictors were included in the parametric component of the PSLISM.

Estimates of parameters obtained from the partially linear single index survival model and AFT-Weibull model for the two data sets are presented in Tables 2 and 3. Table 3 shows that the estimates of the parameters in the parametric component (β) are similar under the ordinary AFT Weibull linear model and the partially linear single-index survival model. However, the estimates of the nonparametric component (single-index) parameter, α, are quite 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 2. Estimates and standard errors of the parameters obtained from the accelerated failure time (AFT) -Weibull model and the partially linear single index model for the culling data set

		AFT - Weibull		PLSISM	
		Estimate	SE	Estimate	SE
AFC	α_1	0.028018	0.01420	0.912131	0.23516
BCS	α_2	0.104516	0.00720	0.249126	0.10294
FAT	α_3	0.001553	0.00052	0.323143	0.14285
HYS	β_1	0.000130	0.00005	0.000310	0.00011
HSV	β_2	0.369476	0.01560	0.296720	0.13701
SUP	β_3	–0.174364	0.02250	–0.231450	0.01352
MILK	β_4	–0.000186	0.00005	–0.002780	0.00143
PROT	β_5	0.009755	0.00056	0.006256	0.00217
Sire	β_6	0.000128	0.00240	0.000203	0.00001

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

For the culling data, Fig. 7 shows that at the beginning as the index increases, the estimated survival time of cows remain somewhat constant and then increases to a maximum. Once it reaches the maximum, as the index increases, the estimated survival time of cows decreases. In the previous section it was shown that as individual components of the index increased (age at first calving, body condition score and fat production), the estimated survival time 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

Table 3. Estimates and standard errors of the parameters obtained from the accelerated failure time (AFT) - Weibull model and the partially linear single index survival model for the calf survival data.

		AFT Weibull		PLSISM	
		Estimate	SE	Estimate	SE
AWTG	α_1	0.014354	0.0035130	–0.342130	0.21010
TPG	α_2	0.012076	0.0035500	0.555160	0.17830
ADG	α_3	0.516200	0.0572000	0.761140	0.22540
HYS	β_1	0.000232	0.0000362	0.000310	0.00010
CE	β_2	–0.072200	0.0107000	–0.106230	0.04350
WWG	β_3	0.013667	0.0020310	0.015601	0.00139
NDI	β_4	0.169772	0.0400950	0.270890	0.15340
sire	β_5	0.045231	0.0023840	0.038520	0.00212

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

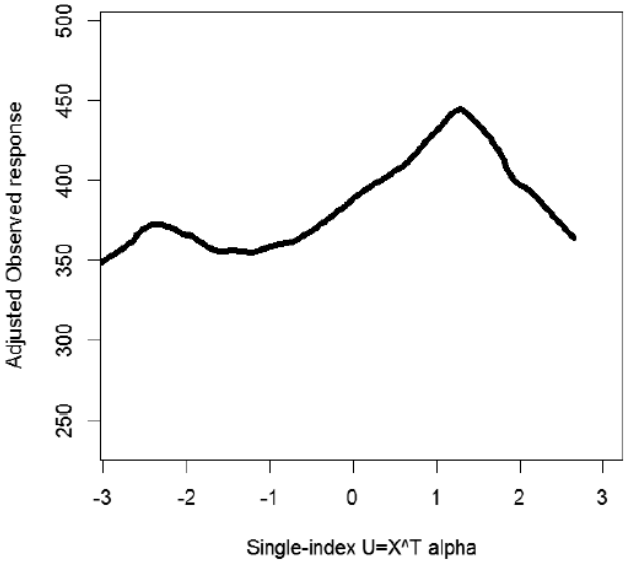

Fig. 7. Observed response against the estimated single index value for the culling data

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 a 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 problems 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 are somehow related and this relationship might affect the survival of cows. Therefore, this relationship may not be explained by the use of a linear parametric survival model.

A similar relationship but a different trend was observed for the calf survival data (Fig. 8). 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, 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 a lower level. This is because the serum protein is important for immune response and ability to resist disease incidence. The serum protein concentration in the blood is a reflection of how well the colostrum feeding and subsequent absorption into 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 regression model.

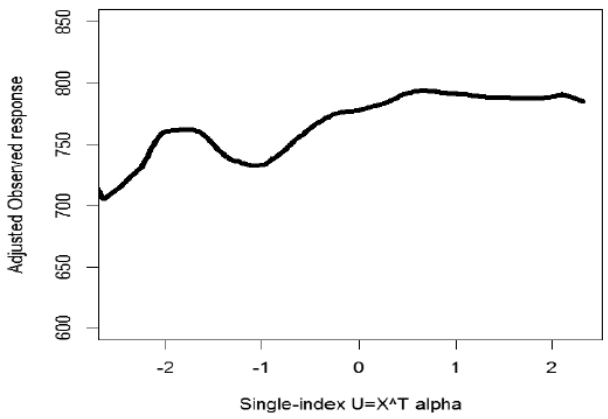

Fig. 8. Observed response against the estimated single index value for the calf survival data.

5. CONCLUSIONS

The present study examined dairy cattle survival data using parametric and semiparametric survival models. The fully parametric model provides nice intuitive interpretations of the results that may have a practical application by providing an estimate of the effect of covariates on survival time. However, in the presence of nonlinearity, these estimated effects might be quite misleading. The semiparametric model reveals nonlinear relationships between survival time and some of the covariates. It is essential to check for, and, if necessary model, such nonlinearities when analyzing biological data. The results have provided some insights, which may be potentially useful in genetic analysis of dairy cattle data, which involves more complex relationship between survival time and covariates. However, further studies should be carried out using larger data sets to validate the current results.

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*. John Wiley & Sons, Inc. New York.

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. Roy. Statist. Soc. B,* **34,** 187-220.

Dabrowska, D.M. (1987). Nonparametric regression with censored survival time data. *Scandinavian J. Stat.*, **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. *Proc. 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. Econ.,* **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 Ann. Statist.*, **25,** 1661-1690.

Fernandez, L. (1986). Non-parametric maximum likelihood estimation of censored regression models. *J. Econ.*, **32,** 35-57.

Henderson, L., Miglior, F., Sewalem, A., Kelton, D. Robinson, A. and Leslie, K.E. (2011). Estimation of genetic parameters for measures of calf survival in a population of Holstein heifer calves from a heifer rearing facility in New York State. *J. Dairy Sci*., **94,** 461-470.

Heuchenne, C. and Van Keilegom, I. (2007). Location estimation in nonparametric regression with censored data. *J. Multi. Anal.,* **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. *Adv. Econ.*, **7**, 45-83.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econ.,* **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.

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

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist.* Assoc., **53,** 457-481.

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. Multi. Anal.,* **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 rightcensored data. *Comm. Statist. —Theory Methods,* **27,** 2895-2907.

Lu, X. and Cheng, T.L. (2007). Randomly censored partially linear single-index models. *J. Mult. Anal.*, **98,** 1985-1922.

Lu, X. and Burke, M.D. (2005). Censored multiple regression by the method of average derivatives. *J. Multi. Anal.,* **95,** 182-205.

Lu, X., Chen, G., Song, X. K. and Singh, R.S. (2006). A class of partially linear single-index survival models. *Canad. 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, O.B. (1995). Kernel estimation in a nonparametric marker dependent hazard model. *The Ann. Stat.,* **23,** 1735-1748.

Powell, J.L. (1986). Censored regression quantiles. *J. Econ.*, **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 Doormaal. 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.

Sewalem, A. (2012). Semiparametric Analysis of Survival Data with Applications in Agricultural Science. M.Sc. Thesis, University of Guelph, Department of Mathematics and Statistics, (http://hdl.handle.net/10214/3650).

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. Appl. Statist.,* **25,** 357-374.

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

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,* **A40,** 945-957.