

Applications in Survival Analysis^{1,2}

STEPHEN D. KACHMAN

Department of Biometry, University of Nebraska-Lincoln, Lincoln 68583-0712

ABSTRACT

Survival or failure time traits such as herd life and days open are both important economically and pose a number of challenges to an analysis based on linear mixed models. The main features of a survival trait are that it is the time until some event occurs, and some of the observations are censored. Survival models and the associated estimation procedures provide a flexible means of modeling survival traits. In this paper I will discuss the application of survival analysis based on the Weibull distribution. The components that make up a survival model will be presented along with their interpretation. Issues related to the model construction and estimation will be presented.

©1999 American Society of Animal Science and American Dairy Science Association. All rights reserved.

(Key words: survival, failure time, mixed model)

INTRODUCTION

Evaluation of traits, which are measured in days, months, or years, poses a number of challenges. These traits consist of the length of time between two events. For example, a breeder may be interested in the length of productive life. The trait would then be the length of time an animal is productive. The breeder is then faced with the following challenges. First, the endpoints of the interval must be defined. Second, how will a record be treated if the animal leaves the herd for a factor unrelated to production? Third, how will a record be treated if the animal is still productive when the evaluation takes place? Fourth, the distribution is heavily skewed. Survival analysis (3, 8) is an approach to analyzing traits such as these.

¹Presented at the 1998 ADSA/ASAS meeting in Denver, Colorado.

²This manuscript has been assigned Journal Series No. 12524, Agricultural Research Division Office.

As its name implies, survival analysis is typically used to examine either the length of time an individual survives or the length of time until a part fails (1, 2, 4, 11). However, survival analysis is also applicable when monitoring length of time until success.

Software such as Proc Lifereg (9) for fixed effects models and, more recently, software such as Survival Kit (5, 6) for mixed models are available to analyze survival data. However, they are not of much use without a basic understanding of survival analysis. In addition, analysis of data sets encountered in animal breeding frequently test the limits of general packages.

The objectives of this paper are to provide a brief introduction to the analysis of survival data and to discuss some of the details needed to modify existing programs to analyze survival traits.

MODEL

The time that animal i fails, T_i , can be thought of as a random process, which depends on many factors. These factors can include fixed effects, β , such as the sex of the animal and random effects, \mathbf{u} , such as the genetic merit of the animal. Typically, these are combined into a vector of risk factors for animal i , $\eta_i = \mathbf{x}_i\beta + \mathbf{z}_i\mathbf{u}$. In an animal breeding analysis the distribution of the random effects is often assumed to be a multivariate normal due to its flexibility in modeling complex covariance structures.

The probability that animal i survives at least until time t , given its risk function, is called the survival function

$$\begin{aligned} S(t; \eta_i) &= \Pr(T_i \geq t) = 1 - F(t; \eta_i) \\ &= \int_t^{\infty} f(w; \eta_i) dw \end{aligned}$$

where T_i = time of failure, $F(t; \eta_i)$ = cumulative distribution function for T_i , and $f(t; \eta_i)$ = density function for T_i .

The challenge is then to develop a reasonable model for the survival function. Hazard functions provide one approach and will be discussed next.

Hazard Function

Models for survival analysis can be built from a hazard function, which measures the risk of failure of an individual at time t . The hazard function for animal i at time t is

$$\lambda(t, \eta_i) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(T_i < t + \Delta t | T_i > t)}{\Delta t} = \frac{f(t, \eta_i)}{S(t, \eta_i)}$$

Another way to look at the hazard function is that for short periods of time (Δt), the probability that an animal fails is approximately equal to $\lambda(t, \eta_i) \Delta t$.

From its definition, the hazard function must be nonnegative. In addition it must be positive at time t unless there is no risk of failure at time t . Without going into detail, the survival function can be obtained from the hazard function with the following relationship

$$S(t, \eta_i) = e^{-\Lambda(t, \eta_i)}$$

where $\Lambda(t, \eta_i) = \int_0^t \lambda(w, \eta_i) dw$. The cumulative distribution and density functions follow directly

$$F(t, \eta_i) = 1 - S(t, \eta_i)$$

$$f(t, \eta_i) = \lambda(t, \eta_i) S(t, \eta_i)$$

If we assume that the risk of failure is constant over time, we get the following hazard function: $\lambda(t, \eta_i) = \lambda$.

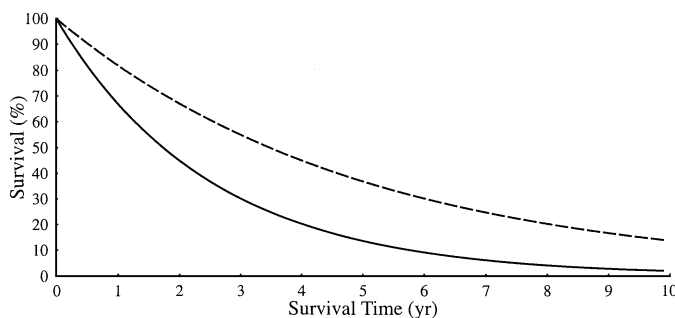


Figure 1. Exponential survival function, $e^{-(\lambda t)}$ where $\lambda = 1/2.5$ (—), or $\lambda = 1/5$ (- -).

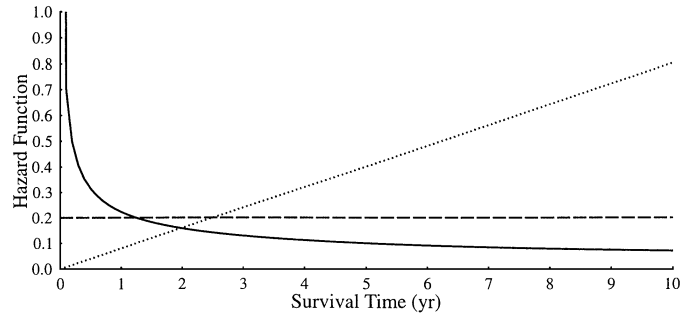


Figure 2. Weibull hazard function, $\rho\lambda(\lambda t)^{\rho-1}$ where $\lambda = 1/5$, and rate parameter (ρ) of 0.5 (—), 1 (- -), or 2 (.....).

$\eta_i) = \lambda$. The resulting density and survival functions are

$$f(t, \eta_i) = \lambda e^{-\lambda t}$$

$$S(t, \eta_i) = e^{-\lambda t}$$

which is an exponential model (Figure 1).

The constant hazard function will produce a population in which the chance of an animal surviving an additional 5 yr is the same at birth and at 5 and 10 yr. A generalization of this would be for the hazard to either increase or decrease over time.

The Weibull model, which has the following hazard function and survival functions

$$\lambda(t, \eta_i) = \rho\lambda(\lambda t)^{\rho-1}$$

$$S(t, \eta_i) = e^{-(\lambda t)^\rho}$$

has the flexibility to model increasing or decreasing hazards. When $\rho = 1$ the Weibull distribution reduces to the exponential distribution. The Weibull model has a decreasing hazard function when $\rho < 1$ and an increasing hazard function when $\rho > 1$. The Weibull hazard and survival functions are presented in Figures 2 and 3.

As can be seen from Figure 3, for a given λ the survival functions all intersect at $t = 1/\lambda$. At $t = 1/\lambda$ the percentage survival is equal to $\exp(-1) \approx 37\%$. The rate parameter ρ determines how quickly the survival function drops off. For small values of the rate parameter, the survival function drops quickly and then levels off. For large values of the rate parameter, the survival function starts off fairly level and then drops off suddenly. The effect of changes in the rate parameter can be seen in Figure 4.

With the Weibull model, λ plays the role of adjusting the intercept. To emphasize the intercept role, the Weibull survival function can be rewritten as

$$S(t; \eta_j) = e^{-\exp[\rho \ln(t) + \rho \ln(\lambda)]} = e^{-\exp[\rho \ln(t) + \eta]} = e^{-t^\rho e^\eta} \tag{1}$$

where $\eta = \rho \ln(\lambda)$. The corresponding hazard function is $\lambda(t; \eta_j) = \rho t^{\rho-1} e^\eta$.

The hazard function is now the product of two parts. The baseline part, $\lambda_0(t) = \rho t^{\rho-1}$, models the basic shape of the hazard function and, therefore, the shape of the density and survival functions. The scaling part, e^η , models the relative risk above or below the baseline risk.

In many cases it is reasonable to assume that the basic shape remains constant for different risk factors but that certain risk factors either increase ($\eta > 0$) or decrease ($\eta < 0$) the overall risk of failure. Proportional hazard models are distributions for which hazard functions can be factored into a baseline hazard, which does not depend on the risk factors, and a scaling factor, which does not depend on time. Typically, the scaling function is $\exp(\eta_j)$ with η_j being a scalar. The hazard function can then be written as

$$\lambda(t; \eta_j) = \lambda_0(t) e^{\eta_j}$$

where $\lambda_0(t) = \lambda(t; 0)$ = baseline hazard function. The survival function can then be written as

$$S(t; \eta_j) = e^{-\Lambda_0(t) e^{\eta_j}}$$

where $\Lambda_0(t) = \Lambda(t; 0)$. The role of the risk factor η_j will be examined next.

Risk Factor η_j

The vector of risk factors is a linear combination of fixed and random effects

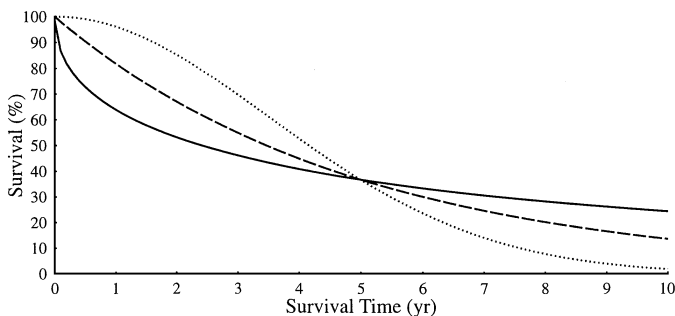


Figure 3. Weibull survival function, $e^{-(\lambda t)^\rho}$ where $\lambda = 1/5$ and rate parameter (ρ) of 0.5 (—), 1 (---), or 2 (.....).

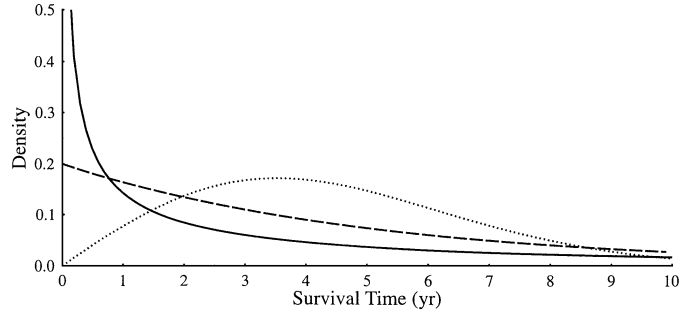


Figure 4. Weibull density function where $\lambda = 1/5$ and rate parameter (ρ) of 0.5 (—), 1 (---), or 2 (.....).

$$\eta = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$$

where β = vector of fixed effects, $\mathbf{u} \sim N(\mathbf{0}, G)$ = vector of random effects, and \mathbf{X} and \mathbf{Z} = known design matrices. The vector of risk factors differs from the usual mixed model equation

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

in several important ways. First, it does not include a residual component (\mathbf{e}). In the survival model the residual variability is modeled through the survival distribution. Second, the expected survival time, given the random effects, is not equal to the $\mathbf{X}\beta + \mathbf{Z}\mathbf{u}$ as in the mixed model. Third, larger values of the risk factor lead to shorter expected survival times.

Under the Weibull survival model, the survival function [1] for animal i can be written as

$$S(t; \eta_j) = e^{-\exp[\rho \ln(t) + \eta_j]} \tag{2}$$

The effect of changes on median survival time, m_{η_j} , can be found by solving $S(m_{\eta_j}; \eta_j) = 0.5$. After a bit of algebra

$$m_{\eta_j} = [-\ln(0.5)]^{1/\rho} e^{-\eta_j/\rho}$$

The effect of a Δ unit change in η_j on median survival time is

$$m_{\eta_j + \Delta} = m_{\eta_j} e^{-\Delta/\rho}$$

For example, let $\rho = 2$, and the risk factor for males be 0.5 larger than the risk factor for females; then the

median survival time for males would be approximately 78% ($e^{-0.5/2} \approx 0.78$) of the median survival time of comparable females. The effect of changes in the risk factor on survival time are in Figure 5.

To summarize, larger risk factors correspond to higher risks and shorter survival times. In addition, an additive change in the risk factor results in a multiplicative change in median survival time.

ESTIMATION

The basic approaches to estimation include non-parametric, semi-parametric, and parametric. The focus of this paper is on the parametric approach. The parametric approach is better suited to handle the large complex models encountered in animal breeding. The basic parametric approach involves getting the joint likelihood of the survival times and the random effects. In simple cases, the marginal likelihood of survival time can be obtained by integrating over random effects. The marginal likelihood can also be approximated by taking a second order Taylor's series expansion of the joint log-likelihood. From a Bayesian viewpoint, that would be equivalent to obtaining the posterior mode.

Ignoring an additive constant, the joint log-likelihood for the Weibull distribution is

$$l(\beta, \mathbf{u}, \rho) = \sum_i [\ln(\rho/t_i) + \rho \ln(t_i) + \eta_i - \exp(\rho \ln(t_i) + \eta_i)] - 1/2 \ln |G| - 1/2 \mathbf{u}' G^{-1} \mathbf{u}$$

Written in a slightly more general form

$$l(\beta, \mathbf{u}, \rho) = \sum_i [\ln(\lambda_0(t_i) + \eta_i - \Lambda_0(t_i) \exp(\eta_i)] - 1/2 \ln |G| - 1/2 \mathbf{u}' G^{-1} \mathbf{u} \tag{3}$$

Posterior mode estimates of the fixed and random effects can then be obtained by taking the first and second partial derivatives of [3]. After a little algebra, the resulting estimation equations are

$$\begin{pmatrix} X'RX & X'RZ \\ Z'RX & Z'RZ + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} X'y^* \\ Z'y^* \end{pmatrix} \tag{4}$$

where

$$\begin{aligned} \mathbf{R} &= - \frac{\partial^2 l}{\partial \eta \partial \eta'} \\ R_{ii} &= \Lambda_0(T_i) e^{\eta_i} \\ y^* &= \frac{\partial l}{\partial \eta} + R\eta \end{aligned}$$

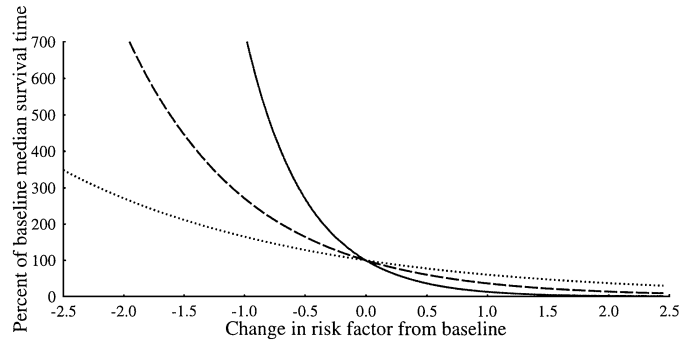


Figure 5. Effect of changes in the risk factor on median survival time with rate parameter (ρ) of 0.5 (—), 1 (---), or 2 (.....).

$$y_i^* = 1 - \Lambda_0(T_i) e^{\eta_i} + R_{ii} \eta_i$$

which is very similar to the usual mixed model equations. Because these equations must be solved iteratively, the computational time will be several times greater than for a corresponding linear model when variance components are known. However, if the variance components must be estimated computational times will be similar to the corresponding linear model. Approximate standard errors and tests are obtained as in the linear model case.

CENSORING

In this section the effect of censoring will be discussed. Unlike traits such as yearling weight, data on survival traits are often censored. That is, the survival time may either be known to be greater than a certain amount (right censored), less than a certain amount (left censored), or be within a certain range (double censored). Of the three types of censoring, right censoring is the most common. Right censoring can occur because an animal is removed before failure can be observed or because failure occurred after the end of data collection. Left censoring can occur because an animal failed before data collection began. Double censoring can occur if there is a break in data collection, and an animal fails somewhere in that interval.

In the following examination of censoring, time of censoring and survival time will be assumed to be independent. Attention will also be focused on handling right censoring. Conceptually other types of censoring are handled similarly.

For data that are right censored, the time of censoring is observed instead of the time of failure. Let T_i

= observed time at which an animal has failed or the time at which the record was censored. If a record is uncensored, then the density function of T_i is

$$f(T_i; \eta_i) = \lambda(T_i; \eta_i) e^{-\Lambda(T_i; \eta_i)} = \lambda(T_i; \eta_i) S(T_i; \eta_i).$$

If a record is censored, then the probability mass function of T_i is obtained by integrating $f(t; \eta_i)$ from T_i to ∞ yielding

$$S(T_i; \eta_i) = e^{-\Lambda(T_i; \eta_i)}.$$

The log likelihood for animal i is

$$l_i = W_i \ln(\lambda(T_i; \eta_i)) - \Lambda(T_i; \eta_i)$$

where $W_i = 1$ if a record is uncensored and $= 0$ if a record is censored. The corresponding elements in [4] are

$$R_{ii} = \Lambda_0(T_i) e^{\eta_i} \quad [5]$$

and

$$y_i^* = W_i - \Lambda_0(T_i) e^{\eta_i} + R_{ii} \eta_i \quad [6]$$

Censoring can lead to difficulties in parameter estimation. When animal i is right censored at T_i , the log likelihood is

$$l_i = -\Lambda_0(T_i) e^{\eta_i}$$

taking the partial with respect to η_i yields

$$\frac{\partial l_i}{\partial \eta_i} = -\Lambda_0(T_i) e^{\eta_i}.$$

Because $\Lambda_0(T_i)$ is positive, and e^{η_i} is positive for all values of η_i , the partial is always negative. The implication is that if all the records in a fixed effect group are right censored, then the estimate of the risk factor for that fixed effect group will go to $-\infty$.

TIME-DEPENDENT COVARIATES

Various events in an animal's life can lead to changes in its risk of failure. For example, the underlying risk in a herd can change over time because of

management, disease, or economic forces. A sharp drop in the price of milk would increase a dairy cow's risk of being culled. However, we would not expect the drop to have an impact on an animal prior to the drop taking place. These changes can also be on an individual animal basis for factors such as disease and reproductive status. Figure 6 illustrates a hazard function for an animal who becomes ill at 2 yr and recovers at 3.5 yr. The risk of failure increases during the period of illness and decreases when the animal recovers. These changes in the risk of failure can be modeled using a time dependent covariate.

The record on the animal is broken into three conditionally independent observations: 1) a well animal with a survival time greater than 2, 2) an ill animal with a survival time greater than 3.5 conditioned on survival till time 2 yr, and 3) a recovered animal conditioned on survival until 3.5 yr. The resulting log likelihood for the animal at time $t > 3.5$ yr is

$$\begin{aligned} l_i(t) &= \ln(S(2; \eta_{i0})) \\ &\quad - \ln(S(2; \eta_{i1})) + \ln(S(3.5; \eta_{i1})) \\ &\quad - \ln(S(3.5; \eta_{i2})) + \ln(f(t; \eta_{i2})) \end{aligned}$$

where η_{i0} , η_{i1} , and η_{i2} = risk factors for animal i when it is well, ill, and recovered, respectively. The likelihood is obtained by observing that

$$\begin{aligned} f(t; \eta_i | t > C) &= \frac{f(t; \eta_i)}{S(C; \eta_i)} \\ l_i(t | t > C) &= \ln(f(t; \eta_i)) - \ln(S(C; \eta_i)) \end{aligned}$$

where $f(t; \eta_i | t > C)$ = density of animal i surviving to time t , conditioned on its surviving to time C , and $l_i(t | t > C)$ = corresponding contribution to the log likelihood.

PROGRAMMING ISSUES

Existing mixed model programs can be modified to handle the analysis of a survival trait with relatively small changes. The changes that need to be made include repeatedly building and solving the mixed model equations with updated risk factors. Within the portion of the program that builds the mixed model equations, risk factors, η_i , adjusted weights, R_{ii} , and adjusted dependent variables, y_i^* , need to be calculated for each animal. The adjusted risk factors can be calculated within the main body of the program. The adjusted weights and dependent variables are

best calculated using a link function to ease future modifications.

The basic changes needed within the main body of the program are illustrated below:

```

DO I=1,N
  Read in record
  ETA=0
  DO J=1, NEFF
    ETA=ETA+X*SOL
  END DO
  CALL LINK(Y,R,ETA,W,YSTAR)
  Build LHS and RHS
END DO

```

Loop to read in the N records.

Calculate the risk factor ($\eta_i = \text{ETA}$) based on the solution for the NEFF fixed and random effects (SOL) and the design matrix (\mathbf{X}).

Calculate the weight ($R_{ii} = R$) and adjusted dependent variable ($y_i^* = \text{YSTAR}$) based on the failure time ($T_i = Y$) and censor code ($W_i = W$).

The two additions are the calculation of the risk factor for animal ETA and the addition of a link subroutine LINK().

The basics of the link subroutine, assuming ρ is known, are

```

SUBROUTINE LINK
(Y,R,ETA,W,YSTAR)
REAL*8 Y,R,ETA,W,
LAMBDA,RHO
RHO=1.
lambda=exp(RHO*log(y))
R=LAMBDA*EXP(ETA)
YSTAR=W-LAMBDA*EXP
(ETA)+R*ETA
RETURN
END

```

Known rate parameter ($\rho = \text{RHO}$)

Baseline survival function ($\Lambda_0(T_i) = \text{lambda}$)

Weight ($R_{ii} = R$) [5].

Adjusted dependent variable ($y_i^* = \text{YSTAR}$) [6]

within the link subroutine the one line that depends on the hazard function selected is the calculation of integrated baseline hazard function lambda.

Although the basics are straightforward, the usefulness of the program will depend on several additional details. First, the above modifications depend on the rate parameter RHO being known. In practice it will need to be estimated. Estimation of the rate parameter can be treated as a second trait and esti-

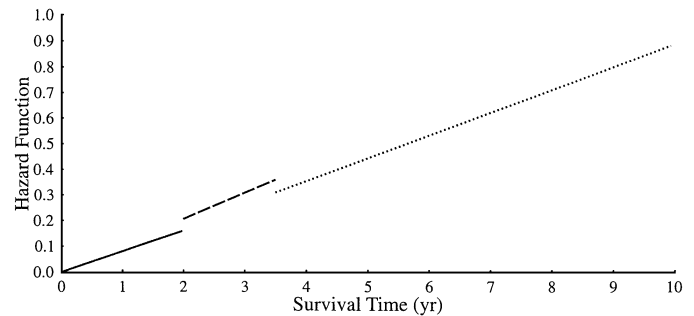


Figure 6. Changes in the hazard function for an animal that starts out well (—), is ill (---) at time 2, and recovers (·····) at time 3.5.

mates obtained by taking partial derivatives of the log likelihood with respect to the rate parameter. The weight matrix and adjusted dependent variables for animal i are

$$R_{ii} = \begin{pmatrix} \Lambda(T_i; \eta_i, \rho_i) \\ \Lambda(T_i; \eta_i, \rho_i) \ln(T_i) \\ \Lambda(T_i; \eta_i, \rho_i) \ln(T_i)^2 + \frac{1}{\rho_i^2} \end{pmatrix}$$

$$y_i^* = \begin{pmatrix} 1 - \Lambda(T_i; \eta_i, \rho_i) \\ (1 - \Lambda(T_i; \eta_i, \rho_i)) \ln(T_i) + \frac{1}{\rho_i} \end{pmatrix} + R_{ii} \begin{pmatrix} \eta_i \\ \rho_i \end{pmatrix}$$

where ρ_i = rate parameter for animal i . Typically the model equation for the rate parameter is $\rho_i = \rho$ or in matrix form

$$\rho = \mathbf{1}\rho$$

where ρ = vector of rate parameters.

In addition, simultaneous estimation of the rate parameter and the risk factor can lead to convergence problems. Often it will be necessary to initially fix the rate parameter to obtain reasonable estimates of the risk factors. Second, risk factors have a tendency to go to $\pm\infty$. Generally this effect is due to contemporary groups in which all observations are right censored or due to inclusion of time-dependent covariates. The basic way of handling this effect is to provide bounds for the risk factors. For example, bounds for $\rho \ln(T_i) + \eta_i$ would be -7 and 2.5 . Third, time-dependent covariates can be handled by preprocessing the data to produce multiple coded records.

SUMMARY

A number of economically important traits measure the time until an event occurs. These traits pose a number of challenges including non-normal distributions and censoring. Survival analysis provides a set of distributions appropriate for time until event traits. In addition, estimation procedures are equipped to handle various forms of censoring. The Weibull survival model with the addition of time-dependent covariates can handle a wide variety of survival traits. Time-dependent covariates also allow the modeling of events which have an effect of limited duration.

In general, modifications to existing programs for evaluating linear mixed models to handle survival traits should be fairly minor. The challenges are primarily in handling boundary conditions and time-dependent covariates. In addition, the Survival Kit set of programs is also available.

This paper provides a brief introduction to survival analysis. A more detailed presentation can be found in Miller et al. (8) which explores various approaches to the analysis of fixed effects models. McCullagh and Nelder (7) provide an introduction from the generalized linear model perspective. Ducrocq and Casella (3) examine the analysis of mixed models from a Bayesian perspective. In addition to the parametric methods discussed in this paper, there are a variety of

nonparametric (8) semi-parametric (10) approaches, which have not been discussed here.

REFERENCES

- 1 Beaudeau, F., V. Ducrocq, C. Fourichon, and H. Seegers. 1995. Effect of disease on length of productive life of French Holstein dairy cows assessed by survival analysis. *J. Dairy Sci.* 78: 103–117.
- 2 Ducrocq, V. 1994. Statistical analysis of length of productive life for dairy cows of the Normande breed. *J. Dairy Sci.* 77:855–866.
- 3 Ducrocq, V., and G. Casella. 1996. A Bayesian analysis of mixed survival analysis. *Genet. Sel. Evol.* 28:505–529.
- 4 Ducrocq, V., R. L. Quaas, E. J. Pollak, and G. Casella. 1988. Length of productive life of dairy cows. 1. Justification of a Weibull model. *J. Dairy Sci.* 71:3061–3070.
- 5 Ducrocq, V., and J. Sölkner. 1998. The Survival Kit—a Fortran package for the analysis of survival data. *Proc. 6th World Cong. Genet. Appl. Livest. Prod.* 23:447–448.
- 6 Ducrocq, V., and J. Sölkner. 1998. The Survival Kit, V3.0, User's Manual. Inst. Natl. Recherche Agron., Universität für Bodenkultur, Jouy-en-Josas, France.
- 7 McCullagh, P., and J. A. Nelder. 1989. *Models for survival data.* Pages 419–431 in *Generalized Linear Models*. 2nd ed. Chapman & Hall, London, United Kingdom.
- 8 Miller, R. G., Jr., G. Gong, and A. Muñoz. 1981. *Survival Analysis.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, NY.
- 9 SAS/STAT® User's Guide, Version 6, 4th Edition. 1989. SAS Inst. Inc., Cary, NC.
- 10 Smith, S. P., and F. R. Allaire. 1986. Analysis of failure times measured on dairy cows: theoretical considerations in animal breeding. *J. Dairy Sci.* 69:217–227.
- 11 Smith, S. P., and R. L. Quaas. 1984. Productive lifespan of bull progeny groups: failure time analysis. *J. Dairy Sci.* 67: 2999–3007.