

GENETIC EVALUATION OF SURVIVAL TRAITS

Stephen D. Kachman

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

Introduction

A survival trait can be broadly defined as the length of time between two events. A typical survival trait is productive life. Length of productive life is the length of time, measured in years, between the start of production until the end of the productive life of the animal. Traits, which are measured in days, months, or years, pose a number of issues which must be addressed before they can be included in a genetic evaluation program. First, the endpoints of the interval must be defined. Second, how will records be handled when the animal leaves the herd? Third, how will incomplete records be handled when evaluation takes place before an animal's second endpoint has been observed? Fourth, how will effects be modeled that occur between the two endpoints? Fifth, how should estimates of the genetic effects be presented to users.

This discussion will describe the use of the Weibull family of survival functions for genetic evaluation. While survival analysis can be used for a number of traits, this guideline will focus on length of productive life.

Definitions

The failure time, T , is the length of time until the animal fails. The survival function, $S(t;\eta)$, is the probability that the animal with a risk level η will survive until time t . The risk level is the sum of effects of factors affecting survival on a log time scale. The median survival, m , is the time at which 50% of animals are expected to fail. In the case of length of productive life, failure would be when a cow is no longer producing.

Hazard Function

Models for survival functions are typically constructed from a hazard function. A hazard function describes the instantaneous failure rate:

$$\lambda(t;\eta) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(T < t + \Delta t | T > t)}{\Delta t}.$$

For short periods of time, Δt , the probability that an animal fails is approximately equal to $\lambda(t;\eta)\Delta t$. Four typical shapes for the hazard function are constant, increasing, decreasing, and bowl shaped. With a constant hazard function, the probability an animal survives an additional year is the same for an animal's first year and fifth year. The resulting model for a constant hazard function is the exponential. The exponential survival function is given by

$$S(t;\eta) = e^{-\exp(\ln(t)+\eta)}$$

which depends on time t , and a risk level η .

The Weibull model is a generalization of the exponential family, which allows the hazard function to either increase or decrease over time. The Weibull model includes an additional rate parameter, ρ . The rate parameter is less than one when the hazard function decreases the longer an animal is in the herd and is greater than one when the hazard function increases the longer an animal is in the herd. The survival function for the Weibull model is given by

$$S(t;\eta) = e^{-\exp(\rho \ln(t) + \eta)}$$

The impact of different rate parameters on survival can be seen in Figure 1. The three curves all have the same median survival time of five years. With a rate parameter of less than one, $\rho = .5$, there are a relatively large number of animals with a very short productive life or a very long productive life. When the rate parameter is greater than one, $\rho = 2$, there are a relatively small number of animals with a very short or a very long productive life.

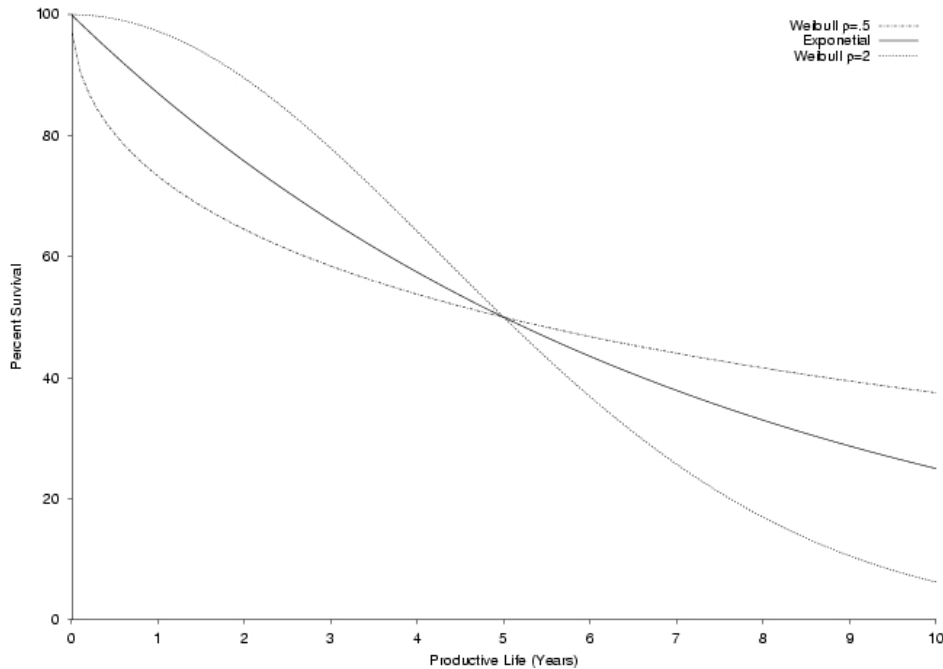


Figure 1. Weibull survival function for three rate parameters and a median productive life of 5 years.

In addition to the Weibull model, there are a number of other models for survival which will not be discussed here.

Risk level

The risk of an animal failing is influenced by genetic, random environmental, and managerial factors. The risk level of an animal can be modeled as the sum of the genetic, environmental, and management effects as

$$\eta = G + E + M,$$

where G is the genetic risk, E is the environmental risk, and M is the contemporary group risk. For the risks to be additive, changes in management factors or environment do not change genetic factors affecting survival. This assumption might not be reasonable, as for example, if culling practices in one herd depend on calf performance and in another herd depend on the reproductive performance of the cow.

Median survival time is one way to gauge the impact of changes in an animal's risk level. For the Weibull distribution, the median survival time for animals at risk level η , is given by

$$m(\eta) = [-\ln(0.5)]^{-1/\rho} e^{-\eta/\rho}.$$

The impact of a change in the genetic risk of Δ on median survival time is given by

$$\frac{m(\eta + \Delta) - m(\eta)}{m(\eta)} = e^{-\Delta/\rho} - 1.$$

The expected percentage changes in median survival time are given in Table 1.

Table 1. Expected changes in median survival time.

| Scaled Genetic Risk (Δ/ρ) | Change in Median Survival (%) | Impact when Median Survival is 5 Years (yr) | Median Survival when Risk is Increased by Δ (yr) |
|---------------------------------------|-------------------------------|---|---|
| -0.5 | 64.9 | +3.25 | 8.25 |
| -0.4 | 49.2 | +2.46 | 7.46 |
| -0.3 | 35.0 | +1.75 | 6.75 |
| -0.2 | 22.1 | +1.11 | 6.11 |
| -0.1 | 10.5 | +0.53 | 5.53 |
| 0 | 0.0 | 0.00 | 5.00 |
| .1 | -9.5 | -0.48 | 4.52 |
| .2 | -18.1 | -0.91 | 4.09 |
| .3 | -25.9 | -1.30 | 3.70 |
| .4 | -33.0 | -1.65 | 3.35 |
| .5 | -39.3 | -1.97 | 3.03 |

Censoring

Unlike traits such as weaning weight, survival traits are often censored. A record is censored for length of productive life when it is known that a cow's length of productive life is at least a certain number of years. Typical causes for a censored record are if the cow was still in production when the final data were collected for analysis or if the animal left the herd for non-production reasons. If censoring is not taken into account,

then the genetic risk for young sires, which could have a large percentage of daughters with censored records, would be biased upwards.

Issues

A number of issues need to be addressed before a survival trait can become an effective part of genetic evaluation program. First, survival must be defined in a biologically meaningful manner. The definition must also make sense across a variety of production environments. Second, selection goals are needed. For example, is the goal to have cows with very long production lives or to have cows that are not likely to be culled early? Third, how should sires with a large number of progeny with censored records be handled? The method described in the Appendix of handling censoring works well when less than 20% of the records are censored. However, young sires are likely to have a large fraction of progeny with censored records.

Appendix: Genetic Evaluation of Survival Traits

Data

Two measurements of survival for a cow need to be recorded: a survival time and a censoring indicator. Survival time (T_i) will be either the actual survival time or the time when the record was censored. The censoring indicator (W_i) will be one if the record is uncensored and zero if it is censored. That is, $W_i = 0$ when the animal is still alive when the record is collected.

Model

Survival can be modeled using a generalized linear mixed model. Two components to be modeled are the risk level (η_i) and the rate parameter (ρ_i) for each animal. The risk level for an animal can be modeled as a linear function of fixed and random effects as with other traits

$$\eta = X_\eta \beta_\eta + Z_\eta u_\eta,$$

where β_η is the vector of fixed effects for risk level (e.g. contemporary groups), $u_\eta \sim N(0, G_\eta)$ is the vector of random effects for risk level (e.g. breeding values), G_η is the covariance matrix for the u_η , and X_η and Z_η are incidence matrices. The rate parameter for an animal can also be modeled as a linear function of fixed and random effects

$$\rho = X_\rho \beta_\rho + Z_\rho u_\rho,$$

where β_ρ is the vector of fixed effects for the rate parameter, $u_\rho \sim N(0, G_\rho)$ is the vector of random effects for the rate parameter, G_ρ is the covariance matrix for u_ρ , and

X_ρ and Z_ρ are incidence matrices. Typically the rate parameter is assumed to be the same for all animals with $X_\rho = 1$ and $Z_\rho = 0$.

The (MAP) maximum a posteriori estimates are obtained by solving

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

with $X = (X_\eta \quad X_\rho)$, $Z = (Z_\eta \quad Z_\rho)$, $\hat{\beta} = \begin{pmatrix} \hat{\beta}_\eta \\ \hat{\beta}_\rho \end{pmatrix}$, $\hat{u} = \begin{pmatrix} \hat{u}_\eta \\ \hat{u}_\rho \end{pmatrix}$, $u \sim N(0, G)$, $R = \begin{pmatrix} R_\eta & R_{\eta\rho} \\ R'_{\eta\rho} & R_\rho \end{pmatrix}$

$$y^* = \begin{pmatrix} y^*_\eta \\ y^*_\rho \end{pmatrix} \quad R_\eta = \text{Diag}(e^{\rho_i \ln(T_i) + \eta_i}), \quad R_\rho = \text{Diag}(e^{\rho_i \ln(T_i) + \eta_i} \ln(T_i)^2) + \frac{1}{\rho_i^2},$$

$R_{\eta,\rho} = \text{Diag}(e^{\rho_i \ln(T_i) + \eta_i} \ln(T_i))$, $y^*_\eta = \{W_i - e^{\rho_i \ln(T_i) + \eta_i}\}_i + R_\eta \eta + R_{\eta,\rho} \rho$, and

$y^*_\rho = \{W_i - e^{\rho_i \ln(T_i) + \eta_i} \ln(T_i)\}_i + R'_{\eta,\rho} \eta + R_\rho \rho$. The η and ρ are the estimates from the previous iteration.

Simultaneous estimation of the risk factor and rate parameter can lead to convergence problems. Frequently such problems can be dealt with by initially fixing the rate parameter until estimates of the risk factor have stabilized. During routine genetic evaluation, the rate parameter can be assumed to be known. Similar to problems encountered with the analysis of threshold traits, estimates of the risk factors can become infinite. One way of handling this is to provide bounds for the risk factors. In practice, bounds for the quantity $\rho \ln(T_i) + \eta_i$ of -7 and 2.5 have worked well.

Additional information

A general introduction to survival analysis can be found in Miller et al. (1981). Recent overviews of the analysis of survival traits from an animal breeding perspective can be found in Ducrocq and Cassella (1996), Kachman (1999), and Vukasinovic (1999).

Ducrocq, V., and G. Cassella. 1996. A Bayesian analysis of mixed survival analysis. *Genet. Sel. Evol.* 28:505-529.

Kachman, S. D. 1999. Applications in Survival analysis. *J. Anim. Sci.* 77, Suppl. 2:147-153.

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.

Vukasinovic, N. 1999. Application of survival analysis in breeding for longevity. In: *Proc. Of the Workshop on Genetic Improvement of Functional Traits in Cattle- Longevity*, May 1999., Jouy-en-Josas, France. *Interbull Bulletin.* 21:3-10