Modeling temporal phenomena in variable environments with parametric models: An application to migrating salmon

Richard W. Zabel [a,*], Brian J. Burke [a], Mary L. Moser [a], Christopher C. Caudill [b]

[a] Northwest Fisheries Science Center, National Marine Fisheries Service, 2725 Montlake Boulevard East, Seattle, WA 98112, United States
[b] Department of Fish and Wildlife, College of Natural Resources, University of Idaho, Moscow, ID 83844, United States

ABSTRACT

Timing phenomena are integral to many ecological processes but are difficult to analyze due to the unique nature of timing data and because environmental conditions and behavior can vary during the observation period. We demonstrated methods, based on parametric hazard-rate modeling, to analyze "time-to-event" data under time-varying conditions. We developed routines in R to apply parametric models, based on the exponential, Weibull, and modified Weibull distributions, to time-to-event data. We applied the models to data on the time for migrating adult salmonids to successfully pass a hydroelectric dam. The model captured pronounced diel behavior and the effects of time-varying covariates river flow, spill, and water temperature on passage times. The methods we demonstrated have potential application to a broad range of ecological questions.

Published by Elsevier B.V.

1. Introduction

The timing and duration of events plays a key role in countless ecological processes. Examples of event times are time to insect visitation at a flower (Muenchow, 1986), duration of tarantula fighting times (Moya-Larano and Wise, 2000), time to establishment for an introduced population (Sahlin et al., 2010), and survival (Anderson, 2000). Typically, event times vary within a population, with the distribution determined by genetic, phenotypic, and environmental variability, and ecological interactions. Further, behavior can vary during the observation period due to varying endogenous and exogenous factors, and this can influence the timing of events. Understanding the mechanisms underlying timing phenomena is paramount to understanding the ecology of timing. However, basic linear regression methods are not well suited for analyzing "time-to-event" data (Hosmer and Lemeshow, 1999), because they typically reduce response and predictor variables to means or medians, thus potentially losing valuable information. In addition, these data are often skewed and can contain "censored" individuals that were not observed to complete the event, which can further complicate analyses.

Time-to-event data are directly analogous to survival time data, however, and epidemiologists have developed powerful methods to analyze this type of data (Cox, 1972; Hosmer and Lemeshow, 1999; Kalbfleish and Prentice, 1980). Survival analysis examines the entire distribution of event times and incorporates censored individuals into the analysis, which eliminates a potential source of bias. Several ecologists have applied survival or "time-to-event" analysis to a range of event types, considering how event times vary in relation to factors such as sex or initial environmental conditions (e.g., Moya-Larano and Wise, 2000; Muenchow, 1986; Sahlin et al., 2010). However, these applications assumed that explanatory covariates, and consequently behavior, remained constant throughout the observation period. In natural populations, conditions vary over time, and this variability can influence the timing of events. Thus another powerful feature of time-to-event analysis is the capability to incorporate time-varying covariates.

In epidemiological studies, the Cox Proportional Hazards (CPH) model has received vast majority applications in survival studies (Carroll, 2003). CPH models are considered "semi-parametric" because the baseline survivorship function is not specified, but the effects of covariates are fully parameterized. The goal of most of these studies is to demonstrate that a particular factor has a proportionate effect on survival, with this effect expressed as an odds ratio. This type of information is important for survival studies, but we contend that ecologists are often more interested in how covariates affect the duration of events, and predicting this response is potentially more precise with a parametric model such as the Weibull model. A further advantage of parametric models is that by comparing alternative forms of the baseline survivorship function in a model comparison analysis, we can get a better understanding of fundamental process that underlie the temporal process of interest.

Because parametric models have not been as popular as CPH models, the use of time varying covariates with parametric models has not been fully supported in many statistical packages, including SAS. We developed routines in R (R Development Core Team, 2008) to fit several parametric survival models to data, and we provide the code as an appendix to this paper. Accordingly, the goal of this paper is to present methods for conducting parametric time-to-event models incorporating time-dependent covariates for ecological field studies.

We apply these methods to migration times of adult Pacific salmonids (*Oncorynchus* spp.). Fish migrating in rivers must frequently overcome barriers, both natural and man-made (Castro-Santos, 2005). In the Columbia River in the northwestern U.S., hydroelectric dams form barriers to upstream migrating fish populations, and the overall impacts of dams are considered an important factor in leading to the decline of salmon populations (National Research Council, 1996) and their subsequent listing under the US Endangered Species Act. Fish ladders are provided for upstream passage, but delays are associated with fish locating these ladders (Caudill et al., 2007; Zabel et al., 2008). Fish can expend considerable energy passing dams, and delays can lead to decreased migration success (Caudill et al., 2007). Therefore, we developed models to relate the time it takes individuals to successfully locate and successfully pass fish ladders to a suite of factors.

2. Methods

Time-to-event modeling. First we define T as a random variable representing times to an event and t_i as the observed time for the ith individual. We designate individuals not observed to complete the event, either because they were lost or did not complete the event by the end of the study, as censored, with c_i representing their last observation time. These "right-censored" individuals contain important information that contributes to the estimation of passage rates (Hosmer and Lemeshow, 1999) and therefore must be included in statistical analyses.

The survivorship function, $S(t)$, is the probability of the event not occurring before t days, or $S(t) = P(T > t)$. To visualize this function, we used the product limit, or Kaplan–Meier method (Hosmer and Lemeshow, 1999; Kalbfleish and Prentice, 1980) using the "survival" package in R (R Development Core Team, 2008), which estimates $S(t)$ based on both observed and censored individuals.

The fundamental feature of time-to-event modeling is the hazard function, $h(t)$, which is the conditional probability that the event will occur during the next short time increment, given that it has not occurred yet (Castro-Santos and Haro, 2003; Ross, 1993):

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \le T < t + \Delta t | T \ge t)}{\Delta t}. \tag{1}$$

The hazard function can vary across time in response to varying environmental conditions or behavior, but most ecological studies have not utilized this capability. A related function, the cumulative hazard function, $H(t)$, determines how much hazard an individual has experienced through time t, and is thus the integration of the hazard function through time t:

$$H(t) = \int_0^t h(\tau) d\tau \tag{2}$$

where τ is a dummy variable for the integration. The survivorship function, $S(t)$, is the probability of the event not occurring before t days, or

$$S(t) = P(T > t) = \exp\ (-H(t)) \tag{3}$$

(Hosmer and Lemeshow, 1999). Note that the survivorship function is simply $1 - F(t)$, where $F(t)$ is the cumulative distribution function

(cdf). Based on the hazard function, the probability density function (pdf) of t, $f(t)$, is $f(t) = h(t) \cdot S(t)$. Thus once the hazard function is specified, all other functions necessary for statistical analyses are derivable from it.

To relate timing events to covariates, a standard assumption is that covariates act multiplicatively on a baseline hazard function, $h_0(t)$:

$$h(t) = h_0(t)\ \exp(\mathbf{x}'\boldsymbol{\beta}) \tag{4}$$

where $\mathbf{x}$ is a vector of covariates, and $\boldsymbol{\beta}$ is a vector of regression coefficients. This is equivalent to assuming that covariates act additively on the log hazard and has the desirable property that the hazard remains positive across all ranges of parameter values. This assumption forms the foundation of Cox Proportional Hazards modeling (CPH, Cox, 1972).

In this analysis, we examined three parametric forms for the baseline hazard function, the exponential, Weibull, and modified Weibull distributions. The exponential model assumes that the baseline hazard is constant through time, or $h_0(t) = \lambda$. The Weibull model is a more flexible model where $h_0(t)$ is specified as $\alpha\lambda^\alpha t^{\alpha-1}$, and thus

$$h(t) = h_0(t)\ \exp(\mathbf{x}'\boldsymbol{\beta}) = \alpha\lambda^\alpha t^{\alpha-1}\ \exp(\mathbf{x}'\boldsymbol{\beta}) \tag{5}$$

If $\alpha = 1$, the baseline hazard function reduces to the exponential function, which will be used here as a null model. If $\alpha < 1$, the hazard function decreases with time (survivorship function exhibits a type III response), and if $\alpha > 1$, it increases with time (survivorship function exhibits a type I response). A drawback of the Weibull distribution is that the baseline hazard increases or decreases monotonically. Therefore, we also tested the modified Weibull distribution (Lai et al., 2003) that can produce "bathtub" (or inverted bathtub) shaped hazard distributions. The hazard function for the modified Weibull distribution is $(\alpha + \gamma t)\lambda^\alpha t^{\alpha-1} \exp(\gamma t)$, which reduces to the regular Weibull distribution when $\gamma = 0$. We note that other parametric baseline hazard models, such as the lognormal, are also commonly used (Allison, 2010; Tableman and Kim, 2004).

Both the CPH and parametric regression models can accommodate time-varying covariates. Here we assume that the covariates are recorded during discrete time intervals, which need not be of equal duration. This is a reasonable assumption that reflects most ecological datasets and makes the calculations below more tractable. To reflect this assumption, we converted Eq. (2) to piecewise integrals:

$$H(t) = \sum_{j=1}^n \int_{z_{j-1}}^{z_j} h(\tau) d\tau = \sum_{j=1}^n [H_0(z_j) - H_0(z_{j-1})] \exp(\mathbf{x}'_j\boldsymbol{\beta}) \tag{6}$$

where the z_js delimit the n discrete time periods, with $z_0 = 0$, and $z_n = t$, $\mathbf{x}_j$ is the vector of covariates during the jth time interval, and H_0 is the cumulative baseline hazard. The summation on the right reflects the accumulation of hazard during discrete intervals.

Likelihood function. The likelihood function is expressed in terms of individuals observed to complete the event (at time t_i) and censored individuals (last observed at time c_i). For the censored individuals, we know that their event time would have been $> c_i$ if they were not censored. Accordingly, for censored individuals, we include $P(T > c_i) = S(c_i)$ into the log-likelihood function:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N_E} f(t_i|\boldsymbol{\theta}) \prod_{i=1}^{N_C} f(c_i|\boldsymbol{\theta}) \tag{7}$$

where N_E is the total number of individuals known to complete the event, N_C is the total number of censored individuals, and $\boldsymbol{\theta}$ is a vector of model parameters that determine the hazard function.