Phanerozoic survival patterns

Peter D Smits
Committee on Evolutionary Biology, University of Chicago

March 10, 2015

1 Introduction

2 Methods

2.1 Fossil occurrence information

Foote and Miller data.

2.2 Survival model

Define y as a vector of length n where the ith element is the duration in geologic stages of genus i, where $i = 1, \ldots, n$.

The simplest parametric suvival model is where species durations are assumed to be distributed exponentially with a single inverse-scale ("rate") parameter λ

$$\begin{aligned} p(y|\lambda) &= \lambda \exp(-\lambda y) \\ y_i &\sim \text{Exponential}(\lambda). \end{aligned}$$

Given a parameteric model of survival, two key functions can be defined: survival $S(t)$ and hazard $h(t)$. $S(t)$ corresponds to the probability that a species having existed for t will not have gone extinct while $h(t)$ corresponds to the instantaneous extinction rate per unit age given taxon age t (**?**). For an exponential model, $S(t)$ is defined

$$S(t) = \exp(-\lambda t) \tag{1}$$

and $h(t)$ is defined

$$h(t) = \lambda \tag{2}$$

An exponential model of duration is the parameteric representation of **?** because the right side of Eq. 2 does not include any t terms.

In order to allow species duration to vary with individual covariate information λ can be reparameterized as a regression (**?**). Given that λ is only defined for all non-negative reals, I use a log-link function. More specifically, this is written out as

$$\lambda = \exp(\beta). \tag{3}$$

β can then be expanded to included covariates of interest.

We can relax assumption of the exponential model that extinction risk is independent of species age by instead assuming that species durations follow a Weibull distribution. The Weibull distribution has a shape parameter α and a scale parameter σ. Note that, by definition, $\sigma = 1/\lambda$. The Weibull distribution and sampling statement wre defined

$$p(y|\alpha, \sigma) = \frac{\alpha}{\sigma}\left(\frac{y}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\frac{y}{\sigma}\right)^{\alpha}\right)$$
$$y \sim \text{Weibull}(\alpha, \sigma).$$

The corresponding $S(t)$ and $h(t)$ functions are defined

$$S(t) = \exp\left(-\left(\frac{t}{\sigma}\right)^{\alpha}\right) \tag{4}$$

$$h(t) = \frac{\alpha}{\sigma}\left(\frac{t}{\sigma}\right)^{\alpha-1}. \tag{5}$$

Similar to the exponential model, to allow for species duration to vary with individual covariate information σ can be reparameterized as a regression (**?**) using a log-link function. α can also be allowed to vary by, for example, hierarchical effects (see below), though it is normally assumed constant for all observations (**?**).

2.2.1 Hierarchical modeling

EXPLAIN THE ADVANTAGES OF HIERARCHICAL MODELING. THE PARTIAL POOLING EFFECT AND HOW THIS IMPROVES THE ESTIMATION OF LOW SAMPLE SIZE GROUPS.

2.2.2 Temporal effects

By definition, a genus originates during a single geologic stage. Define an origination cohort, c_j, as the set of all genera that origination during the same geologic stage j where $j = 1, \ldots, J$, J being the total number of origination cohorts (**?**). Origination cohort is modeled as a hierarchical effect where cohorts are considered exchangeable and drawn from a shared normal distribution. This is written as

$$c_j \sim \mathcal{N}(0, \sigma_c) \tag{6}$$

where the scale parameter σ_c is estimated from the data. This term is incorporated in to Eq. 3 as follows

$$\lambda = \exp(\beta + c_{j[i]}) \tag{7}$$

This approach for handling temporal effects can be further expanded, allowing for more temporal structure to be included. For example, each geologic stage can be considered a member of a specific period of "background extinction," or temporal range between two mass extinction events. Effectively, this means modeling the macroevolutionary regime membership of each cohort (**?**). Define regime r_k as the set of all origination cohorts within the temporal span between two mass extinctions k where $k = 1, \ldots, K$, K being the total number of regimes. Regime is then modeled as a hierarchical one "level" above origination cohort, and is also considered exchangeable and drawn from a shared normal distribution. Equation 6 is then expanded as follows

$$\begin{aligned} c_j &\sim \mathcal{N}(\mu_{k[j]}, \sigma_c) \\ \mu_k &\sim \mathcal{N}(0, \sigma_\mu). \end{aligned} \tag{8}$$

As above, both σ_c and σ_μ are estimated from the data.

The current formulation (Eq. 8) assumes that variance is homogeneous between regimes, or that each regime is considered to have equal variance. To allow for each σ_c to vary by regime, Eq. 8 is further expanded as follows

$$\begin{aligned} c_j &\sim \mathcal{N}(\mu_{k[j]}, \sigma_{k[j]}) \\ \mu_k &\sim \mathcal{N}(0, \sigma_\mu) \\ \sigma_k &\sim \log\mathcal{N}(0, \varsigma_\sigma). \end{aligned} \tag{9}$$

Similar to both previous states, now σ_μ and ς_σ are estimated from the data.

2.2.3 Taxonomic effects

By definition, each genus belongs to a single higher-level taxonomic grouping. Define a higher-level taxonomic grouping l_g as the set of all genera belonging to a shared Linnean classification of approximately class level where $g = 1, \ldots, G$, G being the total number of observed groupings. I model taxonomic groupings as a hierarchical effect where groups are conspired exchangeable and drawn from a shared normal distribution. I consider this approach appropriate because there is no comprehensive phylogenetic hypothesis including all genera from major marine groups across the entire Phanerozoic. In cases where there is a more detailed phylogenetic hypothesis, it would be possible to model phylogenetic effect as an individual hierarchical effect modeled as a multivariate normal distribution with covariance matrix known up to a constant (**??**).

Given the above assumptions, l is modeled as

$$l_g \sim \mathcal{N}(0, \sigma_l) \tag{10}$$

where σ_l is estimated from the data. Just as with temporal effect, this term is incorporated into Eq. 3 as

$$\lambda = \exp(\beta + l_{g[i]}). \tag{11}$$

Incorporating both temporal and taxonomic effects into Eq. 3 as simply

$$\lambda = \exp(\beta + c_{j[i]} + l_{g[i]}). \tag{12}$$

It is possible to increase the complexity of Eq 10 by including increasingly more taxonomic levels both above and below the class level. However, I chose to not do this as it adds what is probably an unnecessary amount of complexity and makes interpretation of results extremely difficult.

What is of interest, however, are Sepkoski's three fauna (**?**) which divide select higher-level classifications into Cambrian, Paleozoic, and Modern fauna (Table 1). In many ways these fauna for the empirical basis of Sepkoski's kinetic model of Phanerozoic diversity (**???**), with each fauna being replaced or out-competed by the subsequent fauna.

To model taxonomic groups as samples from a fauna, first I define a fauna f_h as the set of all taxonomic groups assigned by **?** where $h = 1, \ldots, H$, H bing the number of fauna (3). Fauna are considered exchangeable and drawn from a shared normal distribution. Eq. 10 is then rewritten as

$$\begin{aligned} l_g &\sim \mathcal{N}(\omega_{h[g]}, \sigma_l) \\ \omega_h &\sim \mathcal{N}(0, \sigma_\omega) \end{aligned} \tag{13}$$

where both σ_l and σ_ω are estimated from the data.

Just as with an earlier model of temporal effect (Eq. 8), the current model of taxonomic effect assumes equal variance between fauna (Eq. 13). To relax that assumption, I allow σ_l to vary by fauna. Eq. 13 is then rewritten as

$$\begin{aligned} l_g &\sim \mathcal{N}(\omega_{h[g]}, \sigma_{h[g]}) \\ \omega_h &\sim \mathcal{N}(0, \sigma_\omega) \\ \sigma_h &\sim \log\mathcal{N}(0, \phi_\sigma) \end{aligned} \tag{14}$$

where σ_ω and ϕ_σ are estimated from the data.

2.2.4 Priors

Given the Bayesian framework used here, what remains is the important step of assigning priors probability statements to all estimated parameters. While many of the (hyper)parameters have already been given (hyper)priors, there currently remain a few improper priors.

Fauna	Taxa
Cambrian	Trilobita, Polychaeta, Monoplacophora (Tergomya), Inarticulata (Lingulata)
Paleozoic	Articulata (Rhynchonellata), Crinodea, Ostracoda, Cephalopoda, Anthozoa, Stenolaemata (Cyclocystoidea), Stelleroidea (Asteroidea, Ophiuridea)
Modern	Gastropoda, Bivalvia, Osteichthyes, Malacostraca, Echinoidea, Gymnolaemata, Demospongea, Chondrichthyes

Table 1: Sepkoski's three evolutionary fauna. In parentheses are the taxonomic names used in this study when there was a conflict between Sepkoski's designations and mine.

The intercept term β is given a weakly informative prior $\beta \sim \mathcal{N}(0, 10)$ reflecting that while little information is directly known, the value is most likely not extremely large or small.

The prior terms for the various scale parameters of the hierarchical effects (Eq. 9 and 14) I use a weakly informative half-Cauchy (C^+) distribution in order to better constrain all parameter estimates (**?**). This is especially appropriate in the case of of Eq. 14 as there are only three factors at the fauna level.

$$\begin{aligned} \sigma_\mu &\sim \mathrm{C}^+(2.5) \\ \varsigma_\sigma &\sim \mathrm{C}^+(2.5) \\ \sigma_\omega &\sim \mathrm{C}^+(2.5) \\ \phi_\sigma &\sim \mathrm{C}^+(2.5) \end{aligned} \tag{15}$$

The half-Cauchy distribution is equivalent to a folded t-distribution with 1 degree-of-freedom (**?**).

Finally, the Weibull shape parameter α is also given a weakly informative half-Cauchy prior $\alpha \sim \mathrm{C}^+(2.5)$.

2.3 Parameter estimation

2.4 Posterior predictive checks