Death and taxa: biological, temporal, and historical effects on mammal species duration

Peter D Smits

Committee on Evolutionary Biology, University of Chicago

Abstract

1 Introduction

2 Methods

2.1 Species information

Fossil occurrence information was downloaded from the Paleobiology Database (PBDB; `http://paleodb.org/`). Occurrence, taxonomic, stratigraphic, and biological information was downloaded for all North American mammals. This data set was filtered so that only occurrences identified to the species level, excluding all "sp."-s. All aquatic and volant taxa were also excluded. Additionally, all occurrences without latitude and longitude information were excluded.

Species dietary and locomotor category assignments were done using the assignments in initial the PBDB which were then reassigned into coarser categories (Table 1). This was done to improve interpretability, increase sample size per category, and make these results comparable to previous studies [18, 23].

Fossil occurrences were assigned to 2 My bins ranging through the entire Cenozoic. Taxon duration was measured as the number of bins from the first bin of occurrence to the last bin of occurrence, inclusive.

Species body size estimates were sourced from a large selection of primary literature and compilations, principally the PBDB, PanTHERIA [19], the Neogene Old World Mammal

This study		PBDB categories
Diet	Carnivore	Carnivore
	Herbivore	Browser, folivore, granivore, grazer, herbivore.
	Insectivore	Insectivore.
	Omnivore	Frugivore, omnivore.
Locomotor	Arboreal	Arboreal.
	Ground dwelling	Fossorial, ground dwelling, semifossorial, saltatorial.
	Scansorial	Scansorial.

Table 1: Species trait assignments in this study are a coarser version of the information available in the PBDB. Information was coarsened to improve per category sample size and uniformity and followed this table.

database (Now; `http://www.helsinki.fi/science/now/`), and other large scale data collection efforts [4, 7, 22, 25, 32]. In many cases, species body mass was estimated from anatomical dimensions such as tooth size. These estimates were made using a variety of published regression equations. See Appendix: Data for a complete list of individual sources and equations.

2.1.1 Bioprovince occupancy

For each 2 My time bin, a bipartite biogeographic network was created between species occurrences and spatial units. Spatial units were defined was 2x2 latitude–longitude grid cells from an azimuthal equal-area map projection. In these bipartite networks, taxa can only be linked to localities and *vice versa*. Taxa are not linked to each other, nor are localities linked. Emergent bioprovinces within the biogeographic occurrence network were identified using the map equation [28, 29]. A bioprovince is a set of species–locality connections that are more interconnected within the group than without. This was done for each bin's biogeographic network using the `igraph` package for R [5, 24]. The relative number of bioprovinces occupied per time bin was then determined for each species.

2.1.2 Semi-formal supertree

Because there exists no phylogenetic hypothesis of all Cenozoic fossils mammals from North America, it was necessary to construct a semi-formal supertree. This was done by combining taxonomic information for all the observed species and a few published phylogenies.

The taxonomic information from the PBDB served as the basis for additional revision. The taxonomy of many species was updated using the Encyclopedia of Life (`http://eol.org/`), which collects and collates taxonomic information in a single database. This was done programatically using the `taxize` package for R [31]. This was additionally correct using

various published phylogenies and taxonomies of fossil mammals [16, 17, 25], producing a tree that was a series of nested polytomies.

Polytomies were resolved resolved with respect to the order of their appearance. The resulting tree was then time scaled using the `paleotree` package via the "minimum branch length" approach with a minimum length of 0.1 My [1]. The minimum length is necessary to avoid zero-length branches which cause the phylogenetic covariance matrix not be positive definite, which is an important convenience for computation (see below). While other time scaling approaches are possible [2, 12] this method was chosen for it's simplicity and not requiring additional information about diversification rates which are of interest in this study.

2.2 Survival model

Species duration were assumed to be drawn from a Wiebull distribution I implemented a Bayesian model of species duration, which was assumed to follow a Weibull distribution (Eq. 1) with shape α and scale σ parameters. σ was defined as an exponentiated regression model whose parameters are defined below (Eq. 2). The model described here was the final model at the end of a continuous model development framework where the sampling and prior distributions were iteratively modified to best reflect theory, knowledge of the data, the inclusion of important covariates, and to fit the data. This follows the approach described in Gelman and Hill [10] and Gelman et al. [11].

$$\begin{aligned} p(y_i|\alpha,\sigma) &= \text{Weibull}(y_i|\alpha,\sigma) \\ &= \frac{\alpha}{\sigma}\left(\frac{y_i}{\sigma}\right)^{\alpha-1}\exp\left(-\left(\frac{y_i}{\sigma}\right)^{\alpha}\right) \qquad (1) \\ \sigma &= \exp\left(\frac{-(h_i+\eta_{j[i]}+\sum\beta^T\mathbf{X}_i)}{\alpha}\right) \qquad (2) \end{aligned}$$

The shape parameter α was assumed constant, as is standard practice in survival analysis, and was given a diffuse half-Cauchy prior.

$\mathbf{X}$ is a $n \times k$ matrix of species level covariates. The covariates of interest included the logit of mean relative occupancy and the logarithm of body size (g). The discrete covariates such as dietary or locomotor category, called index variables, were transformed into $n \times (k-1)$ matrices where each column is an indicator variable (0/1) for that species's category, k being the number of categories of the index variable. Only $k-1$ indicator variables are necessary as the intercept takes on the remaining value. Finally, a vector of 1-s were included in the matrix $\mathbf{X}$ whose corresponding β coefficient is the intercept, making $\mathbf{X}$ have eight columns.

β is a vector of regression coefficients, where each element was given a unique, weakly informative Normally distributed prior. These priors were chosen because it is expected that

the effect size of each variable on duration will be small, as is generally the case with binary covariates. In all cases, posterior inference was not effected by changes to this choice of prior.

Regression coefficients are not directly comparable without first standardizing the input variables to have equal standard deviations. This linear transform greatly improves the interpretability of the coefficients as expected change in mean duration given a difference of one standard deviation of the covariate [30]. However, because the expected standard deviation for a binary variable is 0.5, in order to make comparisons between the binary and continuous variables, the continuous inputs were divided by twice their standard deviation [9]. The above model was fit with both unstandardardized and standardized inputs for illustrative purposes.

2.2.1 Hierarchical effects

The two hierarchical effects of interest in this study are origination cohort and shared evolutionary history, or phylogeny. Hierarchical modeling can be considered an intermediate between complete and no pooling [10], where complete pooling is when the differences between groups are ignored while no pooling is where different groups are analyzed separately. By allowing for partial pooling, we are modeling a compromise between these two extremes which allows for better inference. This is done by having all of the groups sharing the same prior with the variance parameter, σ^2, estimated from the data, which then acts as an indicator of the amount of pooling. $\sigma^2 = 0$ and $\sigma^2 = \infty$ indicate complete and no pooling, respectively. Hierarchical modeling is analogous to mixed-effects modeling [10].

Origination cohort is defined as the group of species which originated during the same 2 My temporal bin. The most recent temporal bin, 0-2 Mya, was excluded, leaving 32 different cohorts. The effect of origination cohort j was modeled with each group being a sample from a common cohort effect, η, which was considered Normally distributed with mean 0, and standard deviation σ_c. The value of σ_c was then estimated from the data itself, corresponding to the amount of pooling in the individual estimates of η_j. This approach is a conceptual unification between paleontological dynamic and cohort survival analysis [3, 6, 26, 27, 36], with σ_c acting as a measure of relative importance of these two end members.

$$\eta_j \sim \mathcal{N}(0, \sigma_c)$$
$$\sigma_c \sim \text{halfCauchy}(0, 2.5)$$

The σ_c was given a half-Cauchy hyperprior following Gelman [8]

The impact of shared evolutionary history, or phylogeny, was modeling as an individual effect, where each observation, i was considered drawn from a multivariate normal distribution, h, who's covariance matrix was assumed known up to a constant, σ_p [14, 21]. More fully, this is written

$$
\begin{aligned}
h &\sim \text{Multi}\mathcal{N}(0, \mathbf{\Sigma}) \\
\mathbf{\Sigma} &= \sigma_p^2 \mathbf{V}_{phy} \\
\sigma_p &\sim \text{halfCauchy}(0, 2.5),
\end{aligned}
$$

with σ_p also given a half-Cauchy hyperprior and $\mathbf{V}_{phy}$ being the phylogenetic covariance matrix. $\mathbf{V}_{phy}$ is defined as an n x n matrix where the diagonal elements are the distance from root to tip, in branch length, for each observation and the off-diagonal elements are the amount of shared history, measured in branch length, between observations i and j.

Both η and h were centered at 0 so that these effects can be interpreted as differences from the intercept of the model.

In total, there are three different variance components in this model: sample variance σ_y^2, cohort variance σ_c^2, and phylogenetic variance σ_p^2. The sample variance σ_y^2 is approximated as the variance of the deviance residuals.

In ordinary linear regression, the residuals are defined as $r = y - y_{est}$. For hierarchical models, this definition is inadequate. For survival analysis, the equivalent values are deviance residuals. To define how deviance residuals are calculated, we first must define the survival function. The survival function $S(t)$ is equivalent to the complementary cumulative density function or tail probability of observing a given value or greater [20]. For a Weibull model, $S(t)$ is defined

$$
S(t) = \exp(-(\sigma * t)^{\alpha}) \tag{3}
$$

where σ is defined as above (Eq. 2).

Given $S(t)$, we can define the cumulative hazard function [20],

$$
\Lambda(t) = -log(S(t)). \tag{4}
$$

Next, we define the martingale residuals, r, which is based in counting theory [35]. Martingale residuals are defined in relation to I, the inclusion vector described below (Sec. 2.2.2)

$$
r_i = I_i - \Lambda(t_i) \tag{5}
$$

Martingale residuals have a mean of 0 and ranges between 1 and $-\infty$ and can be viewed as the difference between the observed number of deaths between 0 and t_i and the expected values based on the model. However, the range of the martingale residuals is difficult to interpret, tend to be asymmetrically distributed, and not equivalent to standard residuals.

The solution to this is to use the deviance residuals, D. This is defined as a function of martingale residuals and takes the form

$$
D_i = \text{sign}(r_i)\sqrt{-2[r_i + I_i log(I_i - r_i)]}. \tag{6}
$$

Finally, σ_y^2 is defined as $var(D)$.

I used ratios of the variances and intraclass correlations to measure the amount of partial pooling. The former, when compared to the sample size of the hierarchical effect such as the number of cohorts, is an estimate of the relative amount of pooling, while the latter is a measure of the relative importance of the different variance components [10]. Phylogenetic heritability, h_p^2 [14], is effectively identical to the intraclass correlation of the phylogenetic effect, $\frac{\sigma_p^2}{\sigma_y^2+\sigma_c^2+\sigma_p^2}$.

2.2.2 Censored observations

An important part of survival analysis is the inclusion of "censored" observations [15, 20] or observations where the failure time has not been observed. The most common censored observation is right censored, where the point of extinction had not yet been observed in the period of study. In this case, this means taxa that are still extant. Left censored observations, on the other hand, correspond to observations that went extinct any time between 0 and some known point. In this study, taxa occurring in only a single time bin were left censored. Because of the minimum resolution of the record, we cannot observe if these taxa went extinct in less than that single bin or not.

There are a number of steps involved with modeling censored data, illustrated here with right censored data. The approach taken here follows Gelman et al. [11] and the stan modeling language manual [34]. Let I be a vector of length n, where $I_i = 1$ means the observation is completely observed and $I_i = 0$ means the observation is censored. Beginning with a simple complete-likelihood equation for when all observations are completely observed

$$p(y|\theta) = \prod_{i=1}^{n} \text{Weibull}(y_i|\theta, 1). \tag{7}$$

The likelihood of the inclusion vector I, in this case, is easily defined as a simple i.i.d. equation

$$\begin{aligned} p(I|y,\theta) &= \prod_{i=1}^{n} p(I_i|y_i,\phi) \\ &= \prod_{i=1}^{n} \begin{cases} 1 & \text{if } (I_i = 1) \text{ or } (I_i = 0 \text{ and } y_i > \phi) \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

where ϕ is the censoring point, the observed species duration.

For valid Bayesian inference, conditioned on the data, the joint likelihood of the fully observed data (y_{obs}) and I is obtained by integrating out the censored data (y_{cen}) from the complete likelihood:

$$
\begin{aligned}
p(y_{obs}, I|\theta, \phi) &= \int p(y, I|theta, \phi) dy_{cen} \\
&= \int p(y|\theta, \phi) p(I|y, \theta, \phi) dy_{cen} \\
&= \prod_{i:I_i=1} \text{Weibull}(y_i|\theta, 1) \prod_{i:I_i=0} \int_{\phi}^{\infty} \text{Weibull}(y_i|\theta, 1) p(I_i|y_i, \phi) dy_i \\
&= \prod_{i:I_i=1} \text{Weibull}(y_i|\theta, 1) \prod_{i:I_i=0} 1 - F_{Weibull}(y_i|\theta, \phi),
\end{aligned}
$$

where $F(x)$ is the cummulative density function which is the equivalent of the probability of observing a value of x or less. $1 - F(x)$ is the complementary cummulative denisty function which is the equivalent of the probability of observing a value of x or greater.

A summary of the entire model, save for calculations for censored observations, along with the exact priors for every estimated parameter is presented in Figure 1.

The parameter posteriors were approximated using a Markov-chain Monte Carlo (MCMC) routine implemented in the Stan programming language [33]. Stan implements a Hamiltonian Monte Carlo using a No-U-Turn sampler [13]. Posterior approximation was done using four parallel MCMC chains. Chain convergence was evaluated using the scale reduction factor, $\hat{R}$. Values of $\hat{R}$ close to 1, or less than or equal to 1.1, indicate approximate convergence. Convergence means that the chains are approximately stationary and the samples are well mixed [11].

Both models with and without phylogenetic effects were estimated. Because inverting a large matrix is a memory intense procedure and because the phylogenetic covariance matrix is only assumed known up to a constant, every iteration of the MCMC would involve solving a very large matrix which is not ideal. In order to speed up the MCMC routine, this aspect of the model had to be reparameterized for efficiency purposes. Because of the size of the covariance matrix a custom multivariate sampler was used (see Appendix: Code).

For the model without phylogenetic effect the four MCMC chains ran for 1000 steps, with the first 500 used as warm-up and the last 500 as samples from the posterior. Because of the added complexity of estimating the phylogenetic effect, all four chains were run 20000 steps thinned to every twentieth sample split evenly between warm-up and sampling.

2.2.3 Posterior predictive checks

The most basic assessment of model fit is that simulated data generated using the fitted model should be similar to the observed. This is the idea behind posterior predictive checks. Using the predictors from each of the observed durations, and randomly drawn parameter estimates

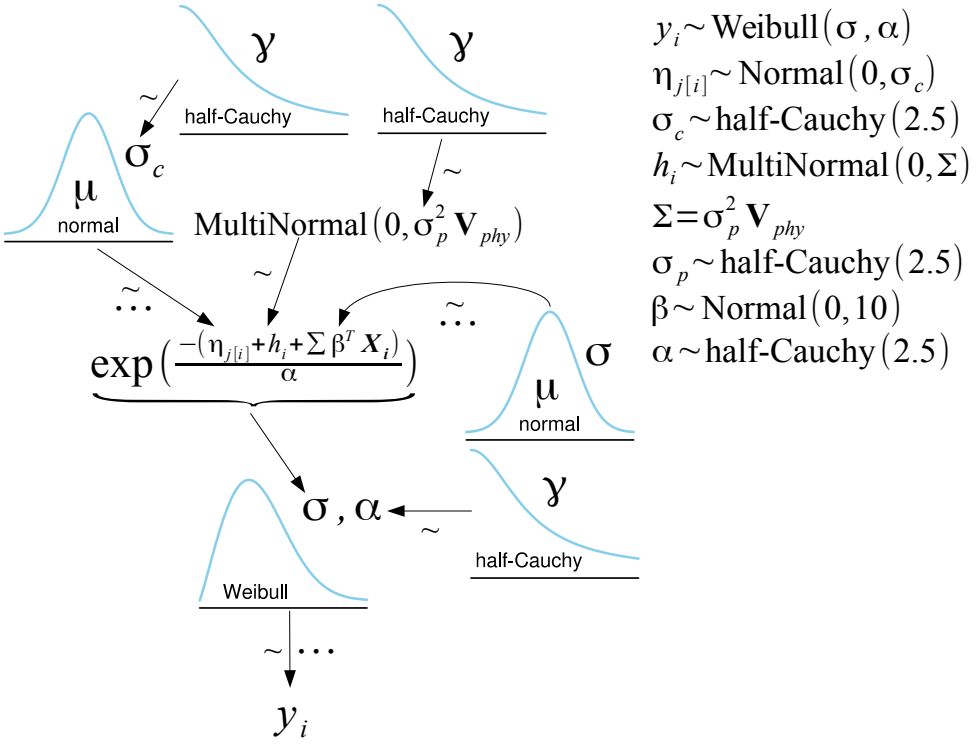

Figure 1: Graphical depiction of full survival model, save censoring, used here. Exact values for the hyperparameters are presented to the right. The observed duration of the ith observation is indicated towards the bottom left as y_i, which is assumed to follow a Weibull distribution.

from their marginal posteriors, a simulated data set y^{rep} was generated. This process repeated 1000 times and the distribution of y^{rep} was compared with the observed y [11]. This was done both graphically and numerically.

An example posterior predictive check used in this study is a graphical comparison of the Kaplan-Meier (K-M) survival curve estimated from the observed data with the survival curves from 1000 simulations. K-M survival curves are non-parametric estimates of the function $S(t)$ or the probability of a species going extinct given that it has lived to time t [20]. Other posterior predictive checks included comparison of the mean and quantiles of the observed durations to the distributions of the same quantities from the simulations.

3 Results

θ	mean	sd	2.5%	25%	50%	75%	97.5%	$\hat{R}$
σ	2.19	0.05	2.11	2.16	2.19	2.23	2.28	1.00
α	1.19	0.03	1.14	1.17	1.19	1.21	1.25	1.00

4 Discussion

References

[1] D. W. Bapst. paleotree: an R package for paleontological and phylogenetic analyses of evolution. *Methods in Ecology and Evolution*, 3:803–807, 2012. doi: 10.1111/j.2041-210X.2012.00223.x. URL `http://doi.wiley.com/10.1111/j.2041-210X.2012.00223.xpapers2://publication/doi/10.1111/j.2041-210X.2012.00223.x`.

[2] D. W. Bapst. A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. *Methods in Ecology and Evolution*, 4(8):724–733, Aug. 2013. ISSN 2041210X. doi: 10.1111/2041-210X.12081. URL `http://doi.wiley.com/10.1111/2041-210X.12081`.

[3] T. K. Baumiller. Survivorship analysis of Paleozoic Crinoidea: effect of filter morphology on evolutionary rates. *Paleobiology*, 19(3):304–321, 1993.

[4] B. W. Brook and D. M. J. S. Bowman. The uncertain blitzkrieg of Pleistocene megafauna. *Journal of Biogeography*, 31(4):517–523, Apr. 2004. ISSN 03050270. doi: 10.1046/j.1365-2699.2003.01028.x. URL `http://doi.wiley.com/10.1046/j.1365-2699.2003.01028.x`.

[5] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.org`.

[6] M. Foote. Survivorship analysis of Cambrian and Ordovician Trilobites. *Paleobiology*, 14 (3):258–271, 1988.

[7] M. Freudenthal and E. Martín-suárez. Estimating body mass of fossil rodents. *Scripta Geologica*, 145:1–130, 2013.

[8] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

[9] A. Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, pages 2865–2873, 2008. doi: 10.1002/sim.

[10] A. Gelman and J. Hill. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY, 2007.

[11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall, Boca Raton, FL, 3 edition, 2013.

[12] M. M. Hedman. Constraints on clade ages from fossil outgroups. *Paleobiology*, 36(1): 16–31, 2010.

[13] M. D. Hoffman and A. Gelman. The no-U-turn sampler: Adaptively setting path lengths

in Hamiltonian Monte Carlo. *arXiv*, 1111(4246), 2011. URL `http://arxiv.org/abs/1111.4246`.

[14] E. A. Housworth, P. Martins, and M. Lynch. The Phylogenetic Mixed Model. *The American Naturalist*, 163(1):84–96, 2004.

[15] J. G. Ibrahim, M.-H. Chen, and D. Sinha. *Bayesian Survival Analysis*. Springer, New York, 2001.

[16] C. M. Janis, K. M. Scott, and L. L. Jacobs. *Evolution of Tertiary mammals of North America. Vol. 1. Terrestrial carnivores, ungulates, and ungulatelike mammals*. Cambridge University Press, Cambridge, 1998.

[17] C. M. Janis, G. F. Gunnell, and M. D. Uhen. *Evolution of Tertiary mammals of North America. Vol. 2. Small mammals, xenarthrans, and marine mammals*. Cambridge University Press, Cambridge, 2008.

[18] J. Jernvall and M. Fortelius. Maintenance of trophic structure in fossil mammal communities: site occupancy and taxon resilience. *American Naturalist*, 164(5):614–624, Nov. 2004. ISSN 1537-5323. doi: 10.1086/424967.

[19] K. E. Jones, J. Bielby, M. Cardillo, S. A. Fritz, J. O'Dell, C. D. L. Orme, K. Safi, W. Sechrest, E. Boakes, C. Carbone, C. Connolly, M. J. Cutts, J. K. Foster, R. Grenyer, M. Habib, C. A. Plaster, S. A. Price, E. A. Rigby, J. Rist, A. Teacher, O. R. P. Bininda-Emonds, J. L. Gittleman, G. M. Mace, and A. Purvis. PanTHERIA : a species-level database of life history , ecology , and geography of extant and recently extinct mammals. *Ecology*, 90(9):2648, 2009.

[20] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2nd edition, 2003.

[21] M. Lynch. Methods for the analysis of comparative data in evolutionary biology. *Evolution*, 45(5):1065–1080, 1991.

[22] R. T. McKenna. *Potential for Speciation in Mammals Following Vast , Late Miocene Volcanic Interruptions in the Pacific Northwest*. Masters, Portland State University, 2011.

[23] S. A. Price, S. S. B. Hopkins, K. K. Smith, and V. L. Roth. Tempo of trophic evolution and its impact on mammalian diversification. *Proceedings of the National Academy of Sciences of the United States of America*, 109(18):7008–12, May 2012. ISSN 1091-6490. doi: 10.1073/pnas.1117133109. URL `http://www.ncbi.nlm.nih.gov/pubmed/22509033`.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

[25] P. Raia, F. Carotenuto, F. Passaro, D. Fulgione, and M. Fortelius. Ecological specialization in fossil mammals explains Cope's rule. *The American naturalist*, 179(3):328–37,

Mar. 2012. ISSN 1537-5323. doi: 10.1086/664081. URL http://www.ncbi.nlm.nih.gov/pubmed/22322221.

[26] D. M. Raup. Taxonomic survivorship curves and Van Valen's Law. *Paleobiology*, 1 (1):82–96, Jan. 1975. ISSN 0036-8075. doi: 10.1126/science.49.1254.50. URL http://www.ncbi.nlm.nih.gov/pubmed/17777225.

[27] D. M. Raup. Cohort Analysis of generic survivorship. *Paleobiology*, 4(1):1–15, 1978.

[28] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–23, Jan. 2008. ISSN 1091-6490. doi: 10.1073/pnas.0706851105. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2234100&tool=pmcentrez&rendertype=abstract.

[29] M. Rosvall, D. Axelsson, and C. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(14):13–24, 2009. URL http://www.springerlink.com/index/H8193132U6432363.pdf.

[30] H. Schielzeth. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1(2):103–113, Feb. 2010. ISSN 2041210X. doi: 10.1111/j.2041-210X.2010.00012.x. URL http://doi.wiley.com/10.1111/j.2041-210X.2010.00012.x.

[31] Scott Chamberlain and Eduard Szocs. taxize - taxonomic search and retrieval in r. *F1000Research*, 2013. URL http://f1000research.com/articles/2-191/v2.

[32] F. A. Smith, J. H. Brown, J. P. Haskell, S. K. Lyons, J. Alroy, E. L. Charnov, T. Dayan, B. J. Enquist, S. K. M. Ernest, E. A. Hadly, K. E. Jones, D. M. Kaufman, P. A. Marquet, B. A. Maurer, K. J. Niklas, W. P. Porter, B. Tiffney, and M. R. Willig. Similarity of Mammalian Body Size across the Taxonomic Hierarchy and across Space and Time. *The American Naturalist*, 163(5):672–691, 2004.

[33] Stan Development Team. Stan: A c++ library for probability and sampling, version 2.5.0, 2014. URL http://mc-stan.org/.

[34] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, Version 2.5.0*, 2014. URL http://mc-stan.org/.

[35] T. M. Therneau, P. M. Grambsch, and T. R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[36] L. Van Valen. Taxonomic survivorship curves. *Evolutionary Theory*, 4:129–142, 1979.