

Article

A Novel Phylogenetic Negative Binomial Regression Model for Count-Dependent Variables

Dwueng-Chwuan Jhwueng *  and Chi-Yu Wu

Department of Statistics, Feng-Chia University, Taichung 407, Taiwan

* Correspondence: dcjhwueng@fcu.edu.tw

Simple Summary: Researchers identified a challenge in analyzing count-dependent variables in species related through a shared ancestry using traditional regression models, as these models often overlook the inherent interdependence from common lineage. To address this, a new phylogenetic negative binomial regression model was developed that recognizes this lineage dependence and allows for overdispersion, surpassing the limitations of the conventional generalized linear models (GLMs). Using the generalized estimating equation (GEE) framework, this model offers precise parameter estimation. This innovation offers a more accurate analysis tool for understanding species data, emphasizing the influence of shared ancestry and promises enhanced research methodologies, bringing valuable perspectives to the fields of evolutionary biology and ecology.

Abstract: Regression models are extensively used to explore the relationship between a dependent variable and its covariates. These models work well when the dependent variable is categorical and the data are supposedly independent, as is the case with generalized linear models (GLMs). However, trait data from related species do not operate under these conditions due to their shared common ancestry, leading to dependence that can be illustrated through a phylogenetic tree. In response to the analytical challenges of count-dependent variables in phylogenetically related species, we have developed a novel phylogenetic negative binomial regression model that allows for overdispersion, a limitation present in the phylogenetic Poisson regression model in the literature. This model overcomes limitations of conventional GLMs, which overlook the inherent dependence arising from shared lineage. Instead, our proposed model acknowledges this factor and uses the generalized estimating equation (GEE) framework for precise parameter estimation. The effectiveness of the proposed model was corroborated by a rigorous simulation study, which, despite the need for careful convergence monitoring, demonstrated its reasonable efficacy. The empirical application of the model to lizard egg-laying count and mammalian litter size data further highlighted its practical relevance. In particular, our results identified negative correlations between increases in egg mass, litter size, ovulation rate, and gestation length with respective yearly counts, while a positive correlation was observed with species lifespan. This study underscores the importance of our proposed model in providing nuanced and accurate analyses of count-dependent variables in related species, highlighting the often overlooked impact of shared ancestry. The model represents a critical advance in research methodologies, opening new avenues for interpretation of related species data in the field.

Keywords: phylogenetic comparative analysis; trait evolution; Poisson regression; negative binomial regression; generalized estimating equation



Citation: Jhwueng, D.-C.; Wu, C.-Y. A Novel Phylogenetic Negative Binomial Regression Model for Count-Dependent Variables. *Biology* **2023**, *12*, 1148. <https://doi.org/10.3390/biology12081148>

Academic Editor: Jacques Demongeot

Received: 9 July 2023

Revised: 16 August 2023

Accepted: 18 August 2023

Published: 19 August 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Phylogenetic comparative methods (PCMs) have a well-established history of illuminating the underpinnings of trait evolution, leveraging the rich insights present within phylogenetic trees [1]. They have traditionally been employed in the analysis of quantitative trait evolution, a practice deeply ingrained in the academic literature [2–6]. Despite

this, an emerging and consistent observation within this area of study is the evolution of categorical traits, which are often represented in categorical or count forms and extend across diverse species. The vast applicability of these methods in a variety of biological scenarios underscores their importance, not only for specialists but also for a broader community of researchers.

Count data has been successfully used to elucidate a range of biological phenomena, for example, the number of toxicological activities in snake venom, or functional activities [7]. In this study, Poisson regression models identified the diversity of diet as a significant predictor of venom's functional activity, demonstrating the value of such models in making ecological predictions. Similarly, in a study on Amazonian forest birds, count data played a crucial role in unraveling the relationship between body mass, flight efficiency, diet, and road-crossing frequency [8]. Here, binomial regression models provided valuable insights into the predictors of road-crossing, which serves as a proxy for the bird's ability to cross habitat gaps—an essential survival skill in the rapidly changing Amazonian landscape. Furthermore, count data in the form of gene copy numbers in yeast species has been utilized to investigate the relationship between metabolic gene copy number and growth rate. A comparative analysis using GEE [9] revealed a clear correlation, providing significant insights into yeast ecology. However, the challenge in such studies often lies in the appropriate analysis of count data. Traditional linear regression forms are ill-suited for such data, since the assumption of normality in the residuals can lead to misleading results when applied to count values.

Hence, there arises the need for alternative models that can adequately account for the specific nature of count data. Enter the realm of GLMs, which includes the phylogenetic Poisson regression [9] and the phylogenetic negative binomial regression which will be developed in this study, that serve as robust tools for such data. Both these models consider the count nature of the data but differ in their assumptions. While the Poisson regression model assumes equal mean and variance, the negative binomial regression model is equipped to handle overdispersion, where the variance exceeds the mean. Although both models find use in different scenarios, it is crucial for practitioners to be aware of potential inaccuracies resulting from the Poisson regression model if the assumption of equal mean and variance is violated [10]. In such scenarios, the phylogenetic negative binomial regression model presents itself as a superior alternative, offering an extra parameter to independently adjust the variance from the mean. This independence can improve model fit and provide more accurate results, highlighting the model's significance.

Furthermore, while the application of the Poisson regression framework is well detailed in previous studies [9], our work focuses on the novel application of the negative binomial regression model in the context of phylogenetic regressions. The remainder of this study, therefore, seeks to introduce this novel phylogenetic negative binomial regression model, test it rigorously, and demonstrate its utility in analyzing count-dependent variables. We believe that the insights gained from this endeavor will provide a fresh perspective to researchers in trait evolution and related fields, enabling a more comprehensive and nuanced understanding of evolutionary dynamics. We will demonstrate the model's efficacy through two distinct empirical assessments: an analysis of lizard egg count as it relates to body mass, and an exploration of mammalian litter size influenced by factors such as the number of teats, longevity, body mass, etc. Through these applications, we hope to underscore the model's utility and contribute to improved methodologies in the study of related species.

The paper is structured as follows: Section 2 outlines our methodology. Section 2.1 discusses regression under a GLM framework, specifically delving into independent Poisson and negative binomial regression. In Section 2.4, we elaborate on the regression under a GEE for phylogenetically dependent data, emphasizing GEE for phylogenetic Poisson and negative binomial regression. Section 3.2 documents our empirical studies on lizard egg-laying and mammalian litter sizes. We present the results of our work, including simulation and empirical analysis outcomes, in Section 3. This is followed by the discussion

in Section 4, and the conclusion in Section 5. Scripts and relevant files developed for this project can be accessed in <https://www.tonyjhwueng.info/phympoinb2reg> accessed on 18 August 2023.

2. Materials and Methods

We present the regression models utilized for analyzing count variables. Traditional linear regression methods are often inadequate for handling count data analysis, primarily due to their assumption of normally distributed residuals, which is unsuitable for count data. In a setting independent of evolution for a group of species, the regression analysis using count data as the response variable and other covariates is analyzed using the GLM described in Section 2.1 where the Poisson regression is described in Section 2.1.1 and the negative binomial regression is described in Section 2.1.2. Note that [11] considered using a single predictor for modeling the count variable under a negative binomial regression model for a couple of empirical data analyzes, our study proposes a general framework concerning multiple covariates and provides a detailed inference. When considering evolution as a dependent process described by a phylogenetic tree relation among species, the regression analysis using count data as response variables and other covariates is carried out by the generalized estimation equation (GEE) in Section 2.4 where the phylogenetic Poisson regression is described in Section 2.4.1 and the phylogenetic negative binomial regression is described in Section 2.4.2.

2.1. Applying GLM in Regression Analysis

GLMs are fundamental tools for regression analysis across various scientific fields, including biology. They offer a flexible statistical framework to analyze different types of response variables, making them an invaluable tool in the biological researcher's toolkit. In the following subsections, we delve into two specific applications of GLMs in the context of biological research: independent Poisson regression and independent negative binomial regression.

The first Section 2.1.1 elaborates on the use of independent Poisson regression, a powerful method particularly suited for the analysis of count data, which is frequently encountered in biological studies. Subsequently, in Section 2.1.2, we turn our attention to negative binomial regression, a model instrumental in handling count data exhibiting overdispersion—a common phenomenon in biological data. Details of the models' mathematical formulation can be found in Appendix A.3, specifically in Appendixes A.3.1 and A.3.2.

2.1.1. Independent Poisson Regression in Biology

Biological research often calls for the analysis of count data—be it bacterial colonies in a dish [12], the number of times a gene gets expressed [13], or species enumerated in an ecological survey [14]. A method conducive to such an analysis is Poisson regression, an efficient instrument to evaluate count data [15]. This technique assumes that the response variable, adheres to a Poisson distribution suitable for count variables, with a mean occurrence rate λ . The probability mass function of the Poisson random variable y is

$$f(y) = \frac{\exp(-\lambda)\lambda^y}{y!}, y = 0, 1, 2, \dots \quad (1)$$

Poisson regression applies a log link function, making it suitable for count data analysis and potentially providing more reliable statistical outcomes [16]. To determine parameters in a Poisson regression model, one can utilize the maximum likelihood estimation (MLE) method, employing numerical strategies such as Newton's method for deriving the MLE [17] (see Appendix A.3.1).

2.1.2. Exploiting Negative Binomial Regression for Overdispersion

In biological studies, researchers frequently confront situations where the response variable is count-based and variable in a way that it surpasses the mean. This phenomenon,

called overdispersion, suggests an inherent data structure that requires careful modeling. In these instances, the negative binomial regression model becomes an instrumental tool for analysis in various biological fields such as the molecular count data from scRNA-seq experiments [18], the weekly dengue haemorrhagic fever cases [19], or the number of fledglings from a nest or inflorescences on a plant [20].

The negative binomial distribution adds an additional parameter (often denoted as r) which models the over-dispersion relative to the Poisson distribution (where the mean equals the variance). This is particularly useful for count data, where often the variance is greater than the mean. The negative binomial model operates under the assumption that the response variable follows a negative binomial distribution. The probability mass function of the negative binomial random variable y is

$$f(y|p, r) = \binom{y+r-1}{y} p^r (1-p)^y, y = r, r+1, r+2, \dots \quad (2)$$

where $0 \leq p \leq 1$ is the probability of success. The model establishes a relationship between the mean response and its predictors through a logarithmic link function, creating a linear relationship with the parameters [17]. This mathematical framework suggests that a systematic alteration in a predictor variable leads to a proportional change in the response. Further details on this can be found in Appendix A.3.2.

In Section 2.2 we provide a preliminary analysis for two empirical datasets using the two count regression models of independent types.

2.2. A Preliminary Analysis

A quick analysis of two empirical datasets using the two GLMs is reported in Table 1 where two fitted regression models (GLM: Poisson regression model vs. the negative binomial model) for the lizard dataset and the mammal dataset are presented. The response variable for the lizard dataset [21] is the egg number per year (EPY) with the covariates egg mass (EM) in gram. The response variable for the mammal dataset [22] is the litter number per year (LY) with another 4 covariates: litter body mass (LS), offspring value as per equation (OV), longevity in years (LG), and whether at least 1 established alien population has successfully spread or not (Spread).

Table 1. Statistical summary and regression under GLM for lizard datasets [21] and mammal [22] datasets. The taxon size n , the mean, and the variance for the response variable $Y = (y_1, y_2, \dots, y_n)$ where each y_i is of count value, the corrected Akaike Information Criterion (AICc) (see Equation (3)) and Akaike weight (w) (see Equation (4)) for each model are reported. NB2 is the abbreviation of the phylogenetic negative binomial model.

Lizard Data		Mammal Data	
Statistics	Value	Statistics	Value
taxon size	17	taxon size	74
mean (Y)	20.82	mean (Y)	1.99
var (Y)	55.03	var (Y)	2.37
Poisson AICc	116.81	Poisson AICc	228.84
NB2 AICc	116.43	NB2 AICc	229.52
Poisson w	0.45	Poisson w	0.58
NB2 w	0.55	NB2 w	0.42

For the mammalian dataset, the variance (2.37) slightly surpasses the mean (1.99), favoring the Poisson regression model, as evidenced by a lower AICc value and a higher weight compared to the negative binomial regression model. In contrast, for the lizard dataset, the variance (55.03) significantly exceeds the mean (20.82) in egg count per year. This discrepancy favors the negative binomial regression model, which has a lower AICc

value and a higher weight compared to the Poisson regression model. This preference for the negative binomial model may be attributed to its unique ability to handle overdispersion, a feature where the phylogenetic negative binomial model particularly excels.

The AICc [23], defined in Equation (3), provides a measure for comparing the quality of different statistical models for a dataset.

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1}. \quad (3)$$

Here, AIC is the Akaike Information Criterion ($2k - 2\ln(\hat{L})$), k is the number of parameters, $\ln(\hat{L})$ is the likelihood value computed from using parameter estimates, and n is the taxa size. The Akaike weights w_i for the i th model measured the importance of the models in the set of candidate models are calculated using Equation (4):

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_{\text{AICc},i}\right)}{\sum_{j=1}^m \exp\left(-\frac{1}{2}\Delta_{\text{AICc},j}\right)}, \quad (4)$$

where $\Delta_{\text{AICc},j} = \text{AICc}_j - \min_{1 \leq j \leq m} \{\text{AICc}_j\}$ [24] represents the difference in AICc values between model j and the model with the smallest AICc value (the best model among m models) and provides a measure of how much worse model j is compared to the best model. Here, $i = 1, 2$ where $i = 1$ for Poisson regression and $i = 2$ for negative binomial regression.

In this equation, $\Delta_{\text{AICc},i}$ is the difference in AICc values between the i th model and the minimal AICc model. The comparison of the fit using the modified Akaike Information criteria (AIC) [25] is shown in Table 1 where the two empirical datasets show a slight preference for either model. For the mammal dataset, the response trait (litter number) has a mean 1.986 and a variance of 2.370. The Poisson regression model provides a slightly better fit to this dataset. For the lizard dataset, the response trait (egg count per year) has a variance of 55.029 and a mean of 20.824. In addition, the regression analysis using covariates: size at maturity, average size, age at maturity, egg mass, clutch size, and clutch mass favors the negative binomial regression model over the Poisson regression model.

In Section 2.3, we introduce the phylogenetic trait evolution of both continuous types as well as the discrete types associated with their count regression model.

2.3. Phylogenetic Trait Evolution

It has been widely accepted that due to speciation and other evolutionary phenomena, species evolved in a dependent manner along a phylogenetic tree. The regression analysis may be more robust when incorporating trees into the analysis. For instance, a five-species phylogenetic tree containing 5 taxa v, u, z, y , and x is presented in Figure 1.

For the continuous trait evolution shown in the lower right panel of Figure 1, trajectories are simulated using the tree traversal algorithm under a continuous random process [26] where five speciation events have occurred in subsequent order, starting at the root ($t = 0$) and continuing immediately afterward. The observed trait values (comparative data) for these five species, represented by v_t, u_t, z_t, y_t , and x_t , are captured at $t = 570$.

The evolution of these traits can be described using the Brownian motion model (BM) [27]. As an example, the trait variable for species v , for example, observed at time t , is expressed as $v_t = \rho + \sigma_v W_t^v$. Here, $\rho = v_0$ denotes the ancestral state of species v , σ_v represents a positive constant parameter, which is the rate of evolution, and W_t^v is a Wiener process, a mathematical construct used in the modeling of stochastic processes. Each species is assumed to have the same rate $\sigma_i = \sigma$, for $i = v, u, z, y, x$ and possess independent identical Wiener processes $W_t^i = W_t$, for $i = v, u, z, y, x$.

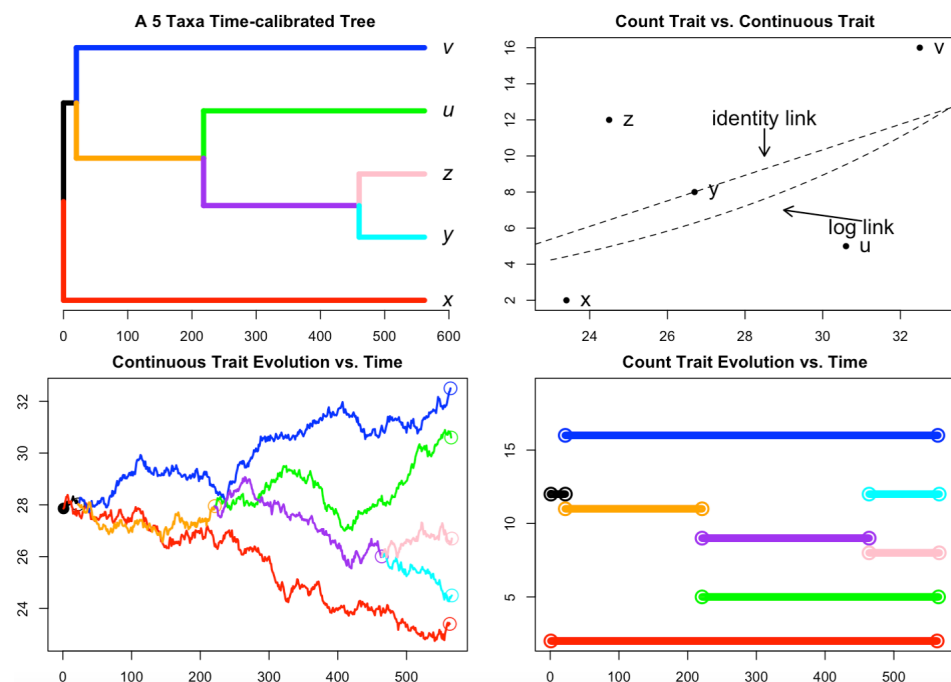


Figure 1. Tree, count trait process, quantitative process, and the bivariate scatter plot. Upper-left panel: a phylogenetic tree of 5 taxa v, u, z, y, x . Upper-right panel: a scatter plot for the count trait data vs. continuous trait data for the 5 taxa. Lower-right panel: a count trait evolution along the phylogenetic tree dependence following a randomized counting process (not necessarily non-decreasing). Lower-left panel: a quantitative trait evolution along the phylogenetic tree with trait values. Each color in the image corresponds to a different branch. The tip values for both the count trait and the continuous are reported in the scatter plot in the upper-right panel. The C matrix corresponding to the tree is shown in Equation (5).

For the count trait evolution shown in the lower-left panel of Figure 1, The tips values at $t = 570$ denoted as (x, y, z, u, v) are assumed to have values $Y = (2, 8, 12, 5, 16)$. Note that one can also consider generating the sample through a tree traversal [28] where starting with the root node with a given value then each successive internal node (the circled points in the figure) is simulated using the status of the starting node plus or minus a Poisson random variable with the rate equal to the branch length multiplied by the status of the nodes where the plus or minus is determined by a Bernoulli trial with value 1 or -1 with probability drawn from a uniform distribution.

It has been known that the tree is incorporated into the analysis for quantitative regression analysis and many packages have been developed to contribute to the community [29–32]. However, conceiving that the negative binomial regression may be potentially useful to analyze count data in phylogenetic regression analysis as the Poisson regression, this work delineates the two phylogenetic regression models for counting dependent variables in a more comprehensive manner using simulation and empirical analysis. In particular, the C matrix will be used for modeling the dependent relationship for the phylogenetic regression using the count response variable. Since the tree can be equivalently transformed into a square matrix C where each element of $c_{ij} \in C$ measures the shared branch length between

the two tips [33,34]. For example, the C for the tree in the upper left panel of Figure 1 can be represented as in Equation (5).

$$C = \begin{matrix} & \begin{matrix} x & y & z & u & v \end{matrix} \\ \begin{matrix} x \\ y \\ z \\ u \\ v \end{matrix} & \begin{pmatrix} 560 & 0 & 0 & 0 & 0 \\ 0 & 560 & 459 & 217 & 20 \\ 0 & 459 & 560 & 217 & 20 \\ 0 & 217 & 217 & 560 & 20 \\ 0 & 20 & 20 & 20 & 560 \end{pmatrix} \end{matrix}. \quad (5)$$

The conceptual regression curves shown in the upper-right panel of Figure 1 using two types of trees and a toy dataset with trait values $Y = (2, 8, 12, 5, 16)$ for dependent count variable, and $x = (23.4, 26.7, 24.5, 30.6, 32.5)$ for quantitative covariate trait variable are shown in Figure 2.

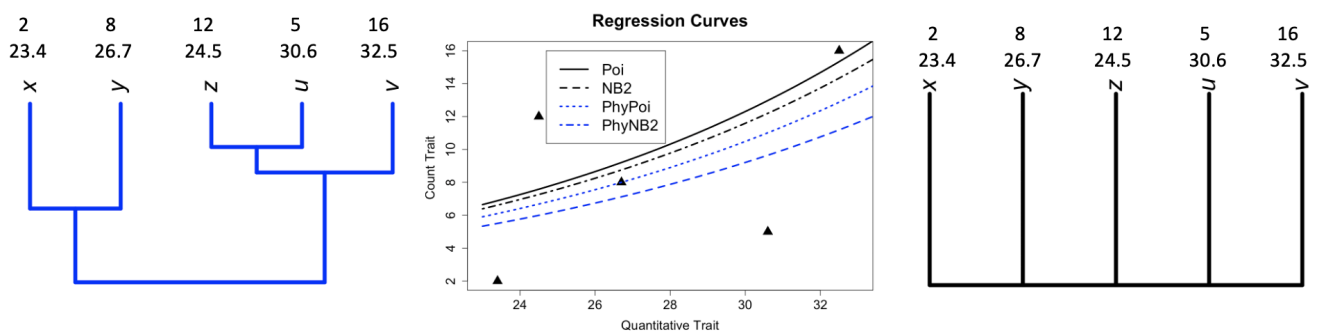


Figure 2. Left panel: phylogenetic tree and two traits. Right panel: star tree and two traits that assume independence. Middle panel: hypothetical regression curves with/without a tree. The regression curves under the GLM approach are reported in black (Poi for Poisson regression and NB2 for negative binomial regression); while the regression curves in blue (PhyPoi for phylogenetic Poisson regression model and PhyNB2 for phylogenetic negative binomial regression model) in blue incorporate the tree (C matrix) under the GEE technique [9] are sketched. The triangles ($\blacktriangle$) are the 2D scatter observation for the response trait (count) vs. predictor trait (quantitative).

2.4. Leveraging GEEs for Regression Analysis of Phylogenetically Dependent Data

Trait evolution research [35], a crucial element in evolutionary biology, requires careful consideration of phylogenetic dependencies embedded within count data. A proven technique to handle these dependencies involves embedding a matrix C , extracted from the phylogenetic tree, in the regression model. This crucial integration accommodates species interrelationships, thereby facilitating precise interpretations. Our analysis primarily focuses on two types of regression models, namely Poisson and negative binomial regression, both members of the exponential family whose probability density function can be expressed in Equation (6) [36].

$$f(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right). \quad (6)$$

GEE emerged as invaluable tools when applying these models. GEE prescribes a parameterization for θ , the distribution parameter of the exponential family, using a link function $g(\cdot)$ that associates the mean function μ and the variance function V of the response variable to the model's linear predictors. Subsequently, the first two moments of y (μ and V), are represented through a series of functional relationships that encompass the parameters θ , μ , η , and $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ where $\eta = g(\mu) = X\beta$ where $X = [x_0, x_1, x_2, \dots, x_p]$ is a design matrix of $n \times (p+1)$ consisting of $x_0 = (1, 1, \dots, 1)^t$ (the vector of 1s) and the covariates $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})$, $j = 1, 2, \dots, p$ [9]. The final estimation equation for

the regression parameter β is obtained by setting the derivative of the $(p + 1)$ estimating equations shown in Equation (7) to zero.

$$\left[\left\{ \sum_{i=1}^n \left(\frac{y_i - \mu_i}{a(\phi)V(\mu_i)} \right) \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \right\}_{j=0,1,2,\dots,p} \right]_{(p+1) \times 1} = [\mathbf{0}]_{(p+1) \times 1}. \quad (7)$$

In the ensuing subsections, we delve deeper into the application of GEE in the domain of phylogenetic trait evolution analysis. We study it in two contexts: the widely acknowledged phylogenetic Poisson regression model and an emerging model, the phylogenetic negative binomial regression model. Given that these regression models are not extensively examined in the current literature, our efforts aim to illuminate their usage and implications, thereby contributing to a broader understanding of phylogenetic trait evolution. Of particular note is the incorporation of the C matrix into the GEE when solving to obtain the estimators (see Equation (9) for Poisson regression case and Equation (11) for negative binomial case). This integration is key to our models where the phylogenetic correlated and dependence among species are used, and the advantages it offers are explicitly discussed in Appendix A, where we lay out the more intricate mathematical details for comprehensive access and understanding. The detailed mathematical formulations of these models are provided in the Appendix A.4, with a specific mention in Appendixes A.4.1 and A.4.2.

2.4.1. Utilizing GEE in Phylogenetic Poisson Regression

Within the domain of evolutionary biology, GEE have become an indispensable tool for scrutinizing count data with inherent correlation structures. This correlation could either be explicitly defined or need estimation. GEE can work with various correlation structures, including independence, exchangeable, autoregressive order 1, and unstructured, as discussed in [15].

A pioneering application of GEE in comparative biology was presented by [9], where the correlation structure is derived from a phylogenetic tree, thereby accounting for the evolutionary interrelations between species. This framework significantly broadens the ability to analyze comparative data, particularly within the Poisson regression model context.

Given a group of n species associated with a trait vector $\mathbf{Y} = (y_1, y_2, \dots, y_n)$. Consider a count response variable y_i for the i th observation with an associated mean rate λ_i . The density function for this variable follows a Poisson distribution and can be represented in an exponential form through a simple logarithmic transformation ($\theta_i = g(\lambda_i) = \log(\lambda_i)$). Within the GEE framework, the first and second moments, $E[y_i]$ and $V[y_i]$, can be derived directly from the link function's derivatives and its inverse Equation (8).

$$\begin{aligned} \mu &= E(y) = b'(\theta) = \lambda, \\ \sigma^2 &= V(y) = b''(\theta)a(\phi) = \lambda. \end{aligned} \quad (8)$$

This approach enables a robust calculation of both the expected value and variance of the response variable, taking into account the phylogenetically structured correlation in the data.

GEE is used to estimate regression parameters in β , employing the chain rule to compute the derivative of the negative log-likelihood function. This process yields an expression involving the i th regression parameter's partial derivative, which can be cast into matrix form, offering a comprehensive perspective on the regression estimates across all observations and parameters. The variance-covariance matrix was further refined [9] for use in phylogenetic comparative analyses, proposing as a combination of the phylogenetic correlation matrix C . The general estimating equation in Equation (7) can be written in matrix form shown in Equation (9).

$$\mathbf{X}_{(p+1) \times n}^t \boldsymbol{\lambda}_{n \times 1} \mathbf{1}_{1 \times n}^t (\mathbf{A}_{n \times n}^{1/2} \mathbf{C}_{n \times n} \mathbf{A}_{n \times n}^{1/2})^{-1} (\mathbf{Y}_{n \times 1} - \boldsymbol{\lambda}_{n \times 1}) = [\mathbf{0}]_{(p+1) \times 1}, \quad (9)$$

where $\lambda = (\lambda_1, \dots, \lambda_n)$ and $A = \text{diag}(\exp(\sum_{j=0}^p \beta_j x_{ij})), i = 1, 2, \dots, n$.

Given a set of response variables y and design matrix X , the regression parameters β can be estimated by solving this nonlinear equation system, providing an exhaustive characterization of trait data within their phylogenetic context (see Appendix A.3.1).

2.4.2. Applying GEE in Negative Binomial Regression

In biological research, the GEE method is in a need of being utilized to perform negative binomial regression. This approach is primarily due to its ability to accommodate overdispersion commonly observed in biological data. It also facilitates adjustments for non-independence resulting from repeated measures, phylogenetic structure, or spatial and temporal autocorrelation, offering significant benefits for applications in evolutionary ecology, population biology, and comparative phylogenetics [37].

In this section, we explore the application of the GEE in negative binomial regression, emphasizing its use in phylogenetic comparative methods. The negative binomial distribution is characterized by parameters r and p , which correspond to the number of successes and the success probability in each trial, respectively.

To conduct a negative binomial regression using the GEE, we employ the canonical log-link function, linking the mean response to the linear predictors. This log-link function, in the context of negative binomial regression, is expressed in terms of r and the mean response μ (i.e., $\theta = \log \frac{\mu}{\mu+r}$). Implementing the GEE necessitates specifying the mean, link, and variance functions. In a negative binomial regression context, the mean function $E[y]$ and the variance function can be written as in Equation (10)

$$\begin{aligned}\mu &= E(y) = r \exp(\theta) / (1 - \exp(\theta)), \\ \sigma^2 &= V(y) = r \exp(\theta) / (1 - \exp(\theta))^2 = \mu + \frac{\mu^2}{r}.\end{aligned}\quad (10)$$

To determine the regression estimates for β_i , we express the link function and the variance function in terms of the observed variables and β_i . Subsequently, we compute the partial derivative of μ_i with respect to η_i , which is crucial for solving the GEE in Equation (7).

From the foundational assumptions, we can derive estimating equations for the regression parameters β . These equations, also referred to as GEE and seen in Equation (7), serve as consistent estimators of β . Their expression in a matrix form, depicted in Equation (11), greatly facilitates solving the nonlinear system for β . In the development of the phylogenetic negative binomial regression, the GEE is transformed into a matrix form to encapsulate the phylogenetic correlation matrix, C . This matrix encodes the phylogenetic relationships among species. The process of integrating C into deriving the phylogenetic negative binomial regression can be represented by the matrix equation in Equation (11).

$$X_{(p+1) \times n}^t \mu_{n \times 1} \mathbf{1}_{1 \times n}^t (B_{n \times n}^{1/2} C_{n \times n} B_{n \times n}^{1/2})^{-1} (Y_{n \times 1} - \mu_{n \times 1}) = [0]_{(p+1) \times 1}, \quad (11)$$

where $\mu = (\mu_1, \dots, \mu_n)$ and $B = \text{diag}(V(y_i | x_i)) = \text{diag}(\mu_i) = \text{diag}(r \exp(\sum_{j=0}^p x_{ij} \beta_j) / (1 - \exp(\sum_{j=0}^p x_{ij} \beta_j)))$. This matrix-based expression of the GEE facilitates solving the nonlinear system for β (see Appendix A.3.2).

The GEE offers a flexible and robust approach to modeling phylogenetic comparative data using negative binomial regression, especially in the presence of overdispersion. Effectively incorporating this into phylogenetic comparative methods can significantly advance our understanding of evolutionary patterns and processes. To test for the significance of the effect, we use the bootstrap technique [38] to generate the samples and re-estimate the parameters for constructing the confidence interval for the empirical analysis. The bootstrap means and the standard error for the regression parameter are reported.

3. Results

To assess the efficacy of our proposed method, we conducted a simulation focused on evaluating the parameter estimation of both regression models. Details regarding the simulation process can be found in Section 3.1. Furthermore, the outcomes specific to the phylogenetic Poisson regression model and the phylogenetic negative binomial regression model are presented in Sections 3.1.1 and 3.1.2, respectively.

3.1. Simulation

To evaluate the method, we performed a simulation to assess the two regression models in the aspect of parameter estimation. The simulation uses four taxa sizes: $n = 16, 32, 64, 128$ and 4 types of trees: coalescent tree, balanced tree, left tree, and star tree. One covariate is used for the assessment of the model and the true parameter for (β_0, β_1) is set to $(3, 5)$. Subsequently, the parameters for simulating responses are computed using the mean function and variance function for the Poisson distribution (as shown in Equation (8)), and the Negative Binomial distribution (as shown in Equation (10)), respectively. The simulation uses 1000 replicates.

Simulate discrete trait: The `ordsample.poi` function we created initiates the generation of simulated data for a phylogenetic Poisson regression model. It produces values from a multivariate normal distribution with zero mean and covariance matrix C derived from the phylogenetic tree. These values are then transformed into Poisson-distributed variables using the `qpois` function, aligning with a Poisson distribution for a particular mean function λ parameter. Consequently, the simulated data mimics count traits with phylogenetic correlation, well-suited for phylogenetic Poisson regression analysis.

Similarly, the `ordsample.nb2` function we created, backed by the MASS library [39], generates simulated data for the phylogenetic negative binomial regression model. It begins by creating random multivariate normal distribution values, consistent with the variance-covariance matrix C of the phylogenetic tree. These values are then transformed into negative binomially distributed variables using the `qnbinom` function with a negative binomial distribution for a particular mean function μ parameter. As a result, the simulated data manifests count traits with phylogenetic dependencies, providing an ideal testing ground for the phylogenetic negative binomial regression model.

When scaling the tree, each branch is assigned a length of less than 1. This can result in zero counts being generated due to the short branch lengths when using count random generators such as a Poisson or negative binomial. Hence, it is imperative to give careful consideration to tree lengths, especially when assessing discrete character changes. Trees of shorter lengths tend to show minimal variation, often exhibiting just 0, 1, or 2 changes from their root to their tip. Hence, expanding these trees by adding more tips might not yield much additional information. Conversely, for elongated trees that average around 15 changes, the varied branches could be more informative, potentially leading to more refined estimates. Instead of merely normalizing tree height, there is merit in exploring the dynamics of taller trees.

Simulate quantitative covariate trait: the predictive trait can be assumed to follow a Brownian motion with root value $\mu = 3$ estimated from the Brownian motion model [40] with rate parameter $\sigma = 1$. This can be directly applied to the multivariate normal distribution $x \sim \mathcal{N}_n(\mu \mathbf{1}, \sigma^2 C)$ as the joint distribution for each Brownian motion random variable is again a normal distribution [33,41]. For non-normal distributed trait, one can consider to simulate the covariate X from the exponential distribution with a known rate parameter.

3.1.1. Phylogenetic Poisson Regression

The response data Y are simulated using the quantile function of the Poisson distribution with the specified mean $\lambda = \exp(\beta_0 + \beta_1 x)$ and the covariate x simulated by the multivariate normal distribution with mean 0 and covariance C . Then, the phylogenetic Poisson regression model is fitted to the samples. For each taxon and tree type

case, 1000 samples are simulated and the mean estimates and standard deviation for the regression parameters are reported in Table 2.

In Table 2, parameter estimates for a phylogenetic Poisson regression model under four types of tree (coalescent, balanced, left, star) and four taxa sizes (16, 32, 64, 128) are presented. Specifically, it reports the mean and standard deviation (in parentheses) of the estimates for the parameters β_0 and β_1 . Furthermore, the means of the parameter estimates seem to be fairly consistent across the various taxa sizes for each tree type. This indicates the robustness of these estimates to the size of taxa considered in the model.

Table 2. The parameter estimates for the phylogenetic Poisson regression. Mean values along with standard deviations are provided for four types of trees across four different taxa sizes.

	Balanced		Coalescent		Left		Star	
	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
16	2.998 (0.086)	5.002 (0.088)	3 (0.034)	5 (0.031)	2.999 (0.059)	5.001 (0.057)	2.999 (0.055)	5 (0.114)
32	3 (0.028)	5 (0.024)	3 (0.009)	5 (0.006)	2.998 (0.082)	5.003 (0.121)	3 (0.029)	5 (0.025)
64	3 (0.025)	5 (0.028)	3 (0.009)	5 (0.008)	3 (0.054)	5.001 (0.071)	3 (0.019)	5 (0.019)
128	3 (0.004)	5 (0.003)	3 (0.008)	5 (0.007)	3 (0.02)	5 (0.019)	3 (0.014)	5 (0.015)

One important observation from the table is the trend of the standard deviations across different taxa sizes, as also shown in Figure 3. For each tree type and parameter (β_0 and β_1), the standard deviation appears to decrease as the taxa size increases from 16 to 128. This suggests that the precision of the parameter estimates improves with increasing taxa size, which is consistent with the idea that larger sample sizes generally provide more precise estimates in statistical analyses. In other words, the estimates for β_0 and β_1 become more reliable and less variable with the increase in taxa size.

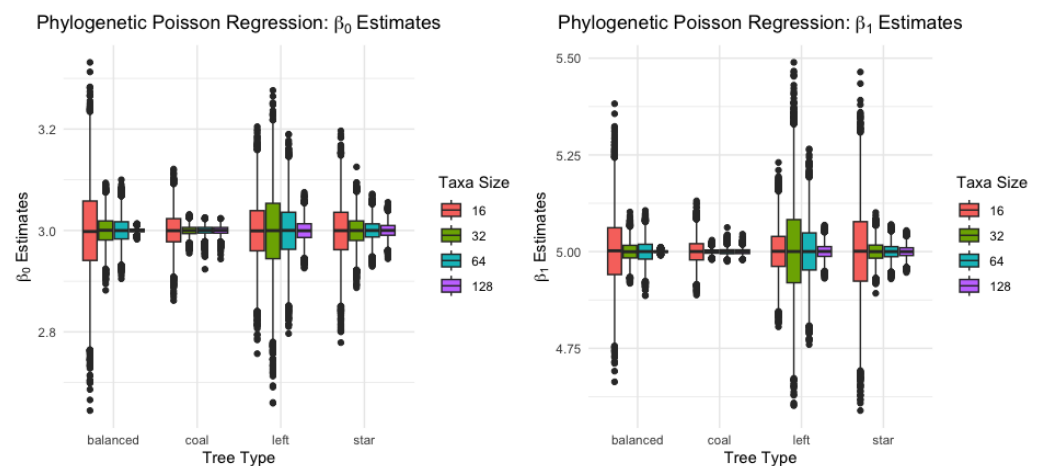


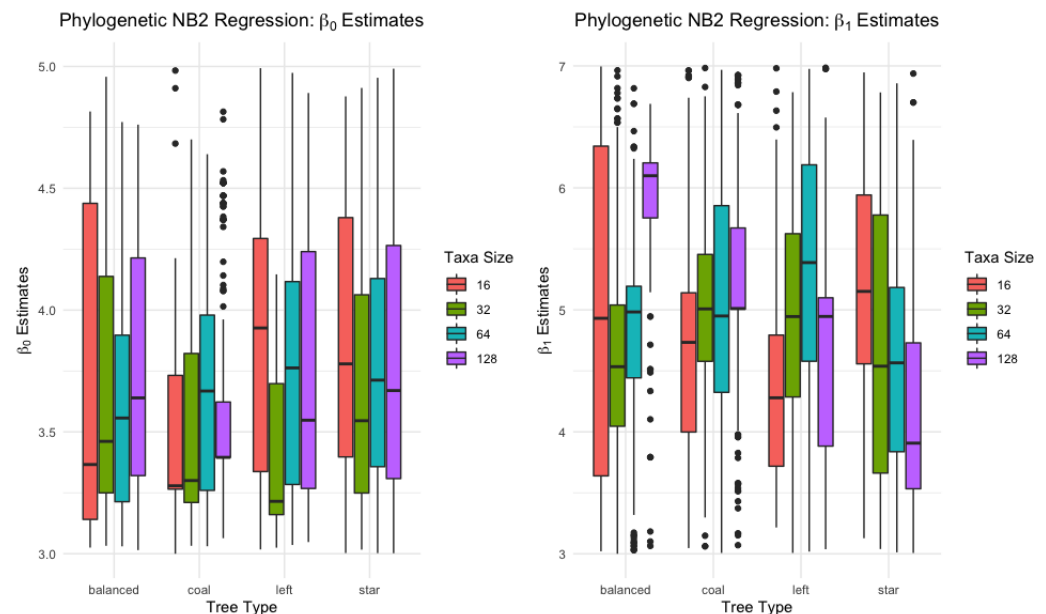
Figure 3. Phylogenetic Poisson regression β_0 and β_1 estimates.

3.1.2. Negative Binomial Regression

Given the covariate samples x , true parameters $\beta_0 = 3, \beta_1 = 5$ and r which is set to 10.68. The response data Y are simulated of the negative binomial distribution with specified mean $\mu = r \exp(\beta_0 + \beta_1 x) / (1 - \exp(\beta_0 + \beta_1 x))$ with dispersion parameter $1/r$. Then, the phylogenetic negative binomial regression model is fitted to the samples. For each taxon and tree type case, 1000 samples are simulated, and the mean estimates and standard deviation for the regression parameters are reported in Table 3 and Figure 4.

Table 3. Parameter estimates for phylogenetic negative binomial regression. The mean and standard deviation under 4 types of tree and 4 taxa size are reported.

	Balanced		Coalescent		Left		Star	
	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
16	1.651 (0.673)	5.029 (1.343)	2.113 (0.864)	4.693 (0.917)	1.878 (0.913)	4.425 (0.929)	2.714 (0.951)	5.226 (1.078)
32	2.029 (0.823)	4.613 (0.924)	2.14 (0.826)	4.97 (0.856)	1.543 (0.548)	4.917 (1.035)	2.705 (0.891)	4.699 (1.117)
64	2.312 (0.831)	4.82 (0.832)	2.031 (0.779)	4.997 (1.008)	1.71 (0.686)	5.255 (1.087)	2.991 (0.899)	4.556 (0.918)
128	2.856 (0.91)	5.856 (0.677)	2.866 (0.802)	5.174 (0.741)	1.79 (0.777)	4.808 (1.052)	3.143 (0.889)	4.21 (1.02)

**Figure 4.** Phylogenetic negative binomial regression β_0 and β_1 estimates.

The parameter estimation results as shown in Table 3 and Figure 4 give valuable insights into the behavior of phylogenetic negative binomial regression across different tree types and taxon sizes.

From the Table 3, it becomes clear that as the taxa size increases, the mean estimates for the intercept (β_0) tend to converge more closely to their true values. Meanwhile, the mean estimates for the slope (β_1) are close to the true value, albeit with a relatively larger standard deviation. This observation reinforces that the phylogenetic negative binomial regression model is performing within expectations, demonstrating its capability to furnish relatively precise parameter estimates across varied conditions. Yet, a deeper exploration into the nuances of parameter estimation within this model reveals challenges in identifying a consistent overarching trend. Some taxa sizes exhibit pronounced variability, marked by significant standard deviations, complicating any straightforward trend interpretation. The quest for consistency across different tree types also proves elusive. This deviation is in sharp contrast to the more discernible patterns typically observed in the phylogenetic Poisson regression model. Such disparities underscore the intricate challenges associated with the phylogenetic negative binomial regression, especially when juxtaposed against other regression frameworks.

One explanation for these larger variations can be found in the nature of the estimation process itself. As mentioned in the text, the estimation of these parameters includes the solving of nonlinear equations (see Equation (11)). Such equations, especially when applied to complex biological data such as phylogenetic trees, can lead to a wide range of solutions. This might explain the relatively large standard deviations observed in these results. It is also worth mentioning that while some variability in the estimates is expected and indeed necessary for the model to adapt to different data structures, overly large variances might

compromise the precision of the model. Therefore, this is a point that might warrant further investigation and potential refinements to the model or the estimation process.

As shown in Table 3 and Figure 4, the high variances could impact the precision of the model. These variances could be a result of the complexity involved in solving nonlinear equations, especially in complex biological data such as phylogenetic trees. Strategies to manage such issues could include employing better algorithms, as will be discussed later, to enhance the solution-finding process. Additionally, lowering the tolerance could help minimize the divergence in results.

By comparing the two models via Tables 2 and 3. Upon comparing the phylogenetic Poisson regression and phylogenetic negative binomial regression models, one notices key differences. The phylogenetic Poisson regression model shows consistent parameter estimates for different taxa sizes, with values for β_0 and β_1 closely clustering around the true values of 3 and 5, respectively, across various tree types. This consistency is accompanied by a remarkably small standard deviation, suggesting a high degree of precision. In contrast, the phylogenetic negative binomial regression model displays more variability in its estimates. Although the values of β_0 and β_1 are in close proximity to the true values, they diverge more than the phylogenetic Poisson regression model's estimates. Additionally, the larger standard deviations point towards greater uncertainty. Despite the higher variability, phylogenetic negative binomial regression could be more suitable under less predictable conditions, while phylogenetic Poisson regression provides stable estimates, proving reliable under steady scenarios.

3.2. Empirical Analysis

Building upon our simulation results, we proceeded to apply our proposed models to real-world empirical datasets. These results served to contextualize and validate our simulated observations, enabling us to examine the models' efficacy in real-life scenarios. The patterns of variability noted in the simulations across tree types and taxa sizes were echoed in the empirical studies, reinforcing our understanding of these dynamics. The use of the phylogenetic negative binomial regression model on the lizard and mammalian datasets also emphasized the model's applicability to count variables in a real biological context. Thus, these empirical analyses provide tangible insights that complement and substantiate our simulation findings.

In our empirical analysis, we currently make use of two different datasets, as outlined in Table 1. The first dataset refers to lizards, with a specific focus on egg count (a count variable) [21]. The second dataset is derived from mammalian data, where the variable of interest is the size of the litter, which refers to the simultaneous live birth of multiple offspring of a single mother [22].

The efficacy of the phylogenetic negative binomial regression model is tested against these two datasets. In Section 3.2.1, we apply this model to the lizard dataset to examine egg count in relation to body mass [21]. For the mammalian dataset, detailed in Section 3.3, we use this model to investigate litter size in response to factors such as number of teats, litter size, longevity, and body mass [22]. These empirical assessments serve to underscore the utility of the phylogenetic negative binomial regression model in the study of count variables.

3.2.1. Lizard's Egg-Laying Count

In various species observed in nature, there appears to be an inverse relationship between egg mass and the number of eggs laid per incubation. For instance, despite having a similar body size to chickens, the kiwi bird produces only one egg, while chickens lay multiple eggs. In our research, we have employed data that were previously collected and studied by [21]. This data primarily focus on the body size, represented as Snout–Vent Length (SVL), of the lizard species *S. undulatus*. Covariates such as age at maturity, egg mass, clutch size, and total eggs were incorporated in the regression analysis, with the response variable being the number of litters.

To enhance the reproducibility of our methodology, we have thoroughly detailed our data pre-processing steps. Initially, the raw data from [21] was collected and compiled in Table A1, found in Appendix A.2.1. This table illustrates the mean values of life history count variables for all *Sceloporus* populations, with the sources for the life-history data and mtDNA specified in the final two columns [42].

We then employed this dataset in our regression analyses, correlating the aforementioned covariates with the number of litters. It is worth mentioning that the phylogenetic tree of the lizard is also based on the study by [21] and is visually represented in Figure 5. The entire process ensures a comprehensive and replicable approach to analyzing the data, thus ensuring the robustness of our findings.

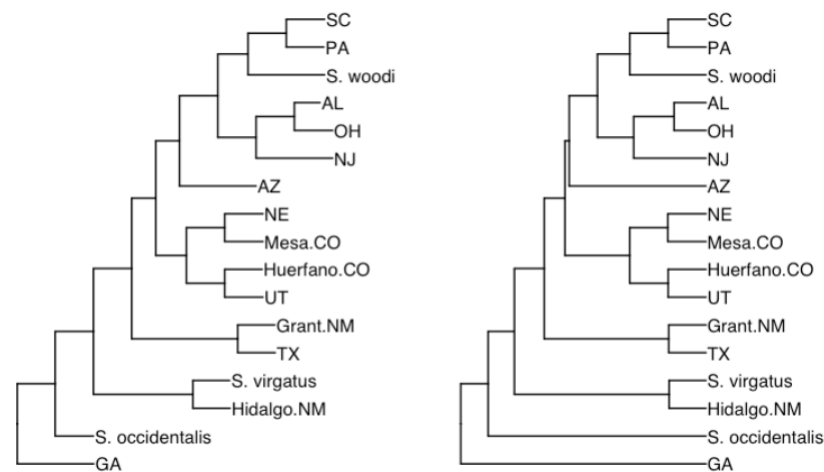


Figure 5. The comparative analysis is rooted in a phylogenetic tree of *Sceloporus undulatus* populations. **Left:** the original tree adapted from a comprehensive tree with branch lengths described by [42] and modified to exclude any populations lacking life-history data. The annotations on the revised tree highlight the state and county of mtDNA sample collection. **Right:** to fit our dataset, this tree is fine-tuned and subsequently transformed it into an ultrametric format using `ape::chronopl` [43]. The visualization of the tree was achieved with `ggtree` [44].

The regression estimates for the model are shown in Table 4.

Table 4. The lizard dataset, sourced from [21], examines the relationship between eggs per year (EPY) and egg mass (EM). It employs bootstrapped estimates and standard deviations, following [38]. Regression methods under a GLM frameworks: `glm.nb` for the negative binomial model and `glm.poi` for the Poisson regression model; and their phylogenetic versions: `phygee.nb` for the phylogenetic negative binomial model and `phygee.poi` for the phylogenetic Poisson model.

	<code>glm.nb</code>	<code>glm.poi</code>	<code>phygee.nb</code>	<code>phygee.poi</code>
Intercept ($\hat{\beta}_0$)	3.397 (0.237)	3.88 (0.172)	3.411 (0.344)	3.753 (0.176)
EM ($\hat{\beta}_1$)	−1.188 (0.735)	−3.302 (0.566)	−1.258 (1.065)	−2.831 (0.569)

Both the Poisson regression coefficient and the negative binomial regression coefficient can be interpreted as follows: for a one-unit change in the predictor variable, the difference in the logs of expected counts of the response variable is expected to change by the respective regression coefficient, given the other predictor variables in the model are held constant.

- In the negative binomial regression (glm.nb), the Egg Mass (EM) coefficient ($\hat{\beta}_1$) is -1.188 . In practical terms, an increase in Egg Mass by one unit results in a decrease in the log of expected counts of Eggs Per Year (EPY) by 1.188 unit. This model, with a standard deviation of 0.735, confirms the inverse association between egg size and the number of eggs laid per year.
- The Poisson regression (glm.poi) exhibits an EM coefficient ($\hat{\beta}_1$) of -3.302 . Meaning, an increase in EM by one unit leads to a decrease in the log of expected counts of EPY by 3.302 unit. With a standard deviation of 0.566, this model reveals a more pronounced inverse relationship between egg size and annual egg production compared to the negative binomial models.
- In the phylogenetic negative binomial regression (phygee.nb), the coefficient of EM ($\hat{\beta}_1$) is -1.258 . This indicates that an increase in EM by one unit results in a 1.258 unit reduction in the log of expected EPY counts. With a standard deviation of 1.065, this phylogenetic model indicates a slightly stronger inverse correlation between egg size and number laid per year than the glm.nb model.
- The phylogenetic Poisson regression via GEEs (phygee.poi) present an EM coefficient ($\hat{\beta}_1$) of -2.831 . This suggests that for every increase in EM of one unit, the log of expected EPY counts decreases by 2.831. The model has a standard deviation of 0.569. Although the phylogenetic model demonstrates a less pronounced effect of egg mass on yearly egg production than the non-phylogenetic Poisson model, it still exhibits a stronger correlation than the negative binomial models.

The comparative analysis of these four models provides some valuable insights. It is noteworthy that the negative binomial models (both general and phylogenetic) show a consistent negative relationship between egg size and annual egg production, albeit with slightly smaller effect sizes. This aligns with existing studies, which also suggest this inverse relationship. However, our work enhances the understanding of this relationship by employing both GLMs and generalized estimation equations, which capture and consider the evolutionary relationship between species.

In comparison, the Poisson models (both non-phylogenetic and phylogenetic) indicate a more pronounced inverse relationship between egg size and annual egg production, which extends the findings of previous research. These results suggest that the use of different statistical models can reveal nuanced details about biological relationships that would not be as evident with a single model. The regression curves are presented in Figure 6.

In summary, the regression models in Table 4 suggest a consistent trend across both negative binomial and Poisson regressions, and their respective phylogenetic versions. All point towards the same biological interpretation: larger egg sizes are associated with fewer eggs being laid per year, with this effect being somewhat stronger in the Poisson models. As illustrated in Figure 6, the negative binomial regression exhibits greater variation and broader confidence intervals than the Poisson regression, whether in phylogenetic or standard contexts. It is worth noting that various genetic and environmental factors can influence egg size in lizards, including the lineage, ambient temperature, and overall health of the animal. A critical observation is the apparent trade-off between egg size and the number of eggs produced annually, potentially representing an adaptive response to optimize offspring survival. Larger eggs might yield stronger, more resilient offspring, but at the cost of reduced egg quantity. This trade-off carries implications for reproductive strategies, population dynamics, and the broader evolutionary course of different lizard species. Understanding this phenomenon further would yield important insights into lizard life history strategies and their responses to environmental changes.

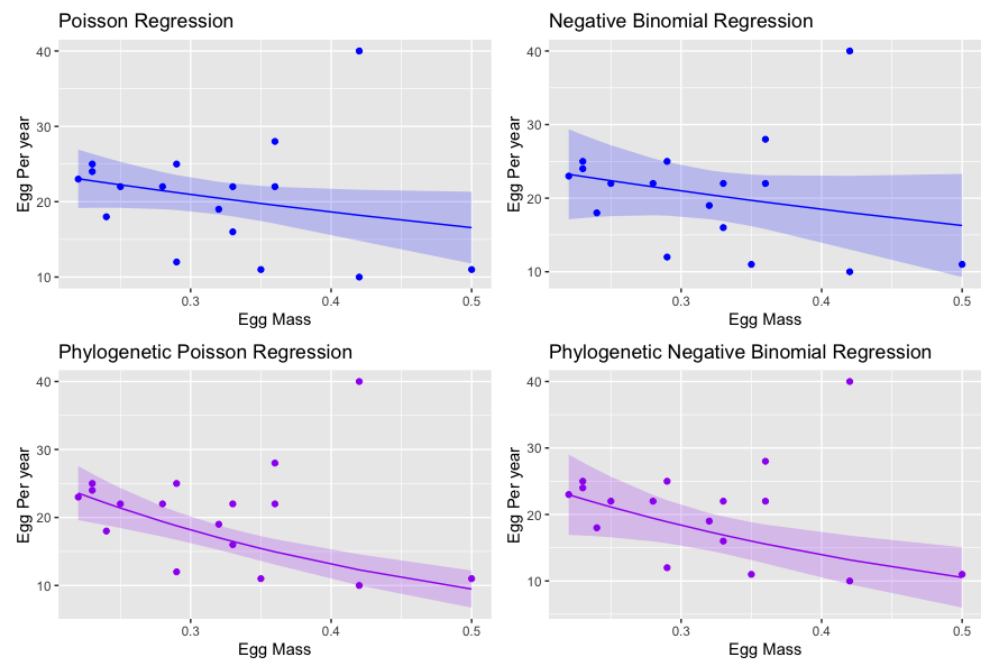


Figure 6. Regression curves for count data: Poisson (**upper left**), negative binomial (**upper right**), phylogenetic Poisson (**lower left**), and phylogenetic negative binomial (**lower right**). Predicted values, $\hat{y}_i$, form the curves. For Poisson-type regressions, $\hat{y}_i$ equals $\exp(\hat{\beta}_0 x_{i0} + \hat{\beta}_1 x_{i1})$, and for negative binomial-type regression, it equals $(\hat{r} \exp(\hat{\beta}_0 x_{i0} + \hat{\beta}_1 x_{i1})) / (1 - \exp(\hat{\beta}_0 x_{i0} + \hat{\beta}_1 x_{i1}))$. The `fitted.values` function retrieves these values for GLM models. For phylogenetic models, the `predict` function follows the usage of `compar.gee` [31] for phylogenetic poisson regression and `compar.gee.nb2` created for phylogenetic negative binomial regression, retrieving the estimates.

3.3. Litter Size in Mammal

In mammals, there is a general pattern where the maximum litter size is often constrained by the number of teats, and typically, the average litter size is about half the number of teats. This trend, however, can vary across different species [45]. For instance, the naked mole-rat (*Heterocephalus glaber*) presents an interesting deviation. It has approximately 12 nipples, but its average litter size is about 11 pups, significantly higher than the typical half. Moreover, the litter size can range from 3 to 12 pups and can even reach as high as 28 in some instances [46].

The need for a comprehensive understanding inspired us to devise a new methodology. Our study incorporates the collection of data pertaining to mammal litter sizes and other traits, such as body mass, gestation length, weaning age, height, and other relevant measurements, as detailed in [22]. The trait data depicted in Table A2 was obtained from [47] (see Appendix A.2.2). We further integrated the mammalian phylogenetic tree, as shown in Figure 7, derived from Phylotastic [48] in a manner similar to [49]. The featured phylogeny encompasses 30 species with complete datasets across all four traits under consideration.

Having discussed the collection and integration of the data, it is crucial to expound on how this gathered data is utilized. This brings us to the application of statistical models, which provide the framework for interpreting the information and yielding insightful findings. Under the assumption that the observations are independently distributed, parameter estimation falls within the purview of the GLM. Progressing to phylogenetic negative binomial regression analysis, initial estimates of parameters are computed using the R package `glm` with the Poisson family. This step solidifies the foundation for our subsequent analysis, ensuring that our data are primed for generating robust conclusions.

The regression estimates for the model are shown in Table 5.

- In the negative binomial regression (`glm.nb`), biological factors impact the expected log count of Litter Size per Year (LY). A one-unit increase in Litter Mean Body Size

(LS) or Offspring Value (OV) reduces the log count of LY by -0.135 and -1.409 unit respectively, all else being equal. Longevity (LG) also has a smaller, negative impact, with a -0.047 decrease per unit increase. Contrastingly, a unit increase in Spread (SP) increases the LY log count by 0.47 unit.

- For the phylogenetic negative binomial regression (phygee.nb), the same biological factors show slightly altered impacts but maintain their directions. The log count of LY decreases by -0.143 , -1.479 , and -0.048 unit with each unit increase in LS, OV, and LG, respectively. However, a unit rise in SP increases the LY log count by 0.478 unit.
- In the Poisson regression (glm.poi), each unit increase in LS, OV, and LG reduces the log count of LY by -0.235 , -2.572 , and -0.058 unit, respectively. Conversely, a unit rise in SP increases the log count of LY by 0.515 unit.
- In the phylogenetic Poisson regression (phygee.poi), each unit increase in LS, OV, and LG leads to a decrease in the log count of LY by -0.231 , -2.621 , and -0.059 unit, respectively. In contrast, a unit rise in SP increases the LY log count by 0.521 unit.

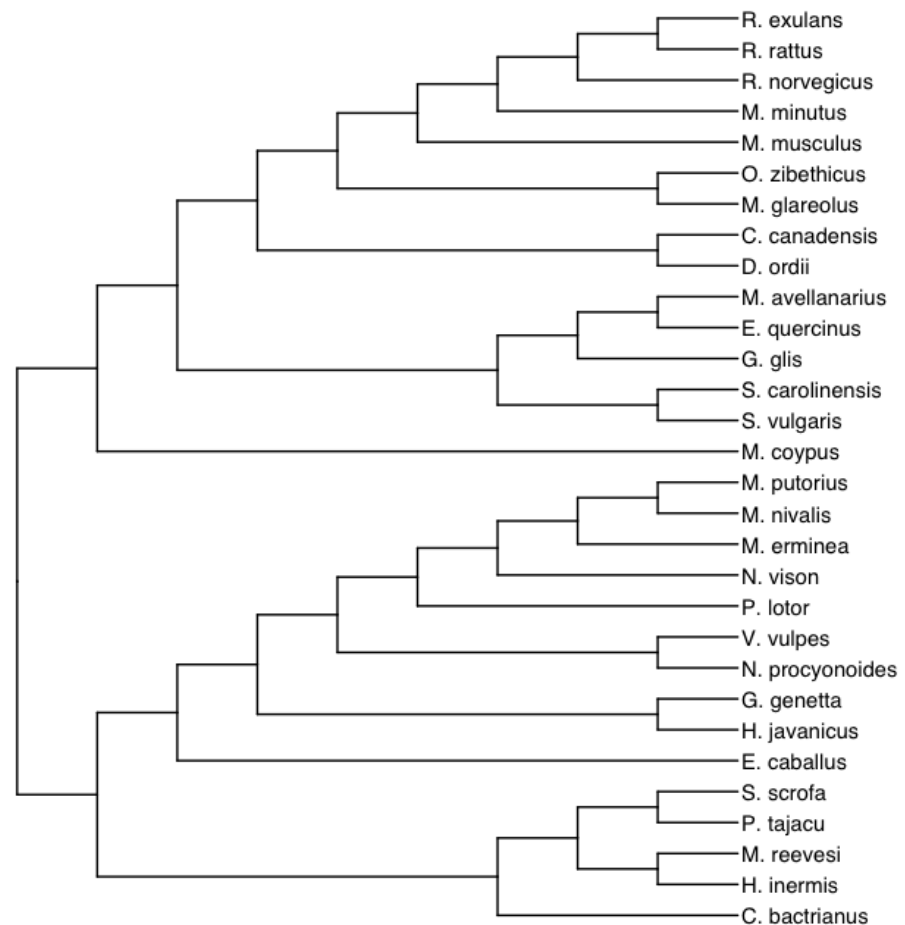


Figure 7. This figure presents a mammalian phylogenetic tree obtained from Phylotastic [48], analogous to [49]. It features 30 species with full data across all four study traits.

In summary, across all models, an increase in each of LS, OV, and LG while holding all other predictors in the model is associated with a decrease in the expected log count of LY, while an increase in SP is associated with an increase in the expected log count of LY. However, the magnitude of these impacts varies between the models. While the Poisson models generally estimate larger effects than the negative binomial models, the negative binomial models accounts for larger variation than the Poisson models. In

addition, the phylogenetic models estimate slightly different impacts compared to their non-phylogenetic counterparts.

Table 5. The mammal dataset, inclusive of bootstrapping estimates and standard deviations, evaluates litter size per year (LY) in relation to litter mean body size (LS), offspring value (OV), longevity (LG), and spread (SP).

	glm.nb	phygee.nb	glm.poi	phygee.poi
Intercept ($\hat{\beta}_0$)	2.13 (0.414)	2.153 (0.414)	2.497 (0.289)	2.492 (0.309)
LS ($\hat{\beta}_1$)	−0.135 (0.105)	−0.143 (0.102)	−0.235 (0.105)	−0.231 (0.106)
OV ($\hat{\beta}_2$)	−1.409 (1.458)	−1.479 (1.616)	−2.572 (0.815)	−2.621 (1.041)
LG ($\hat{\beta}_3$)	−0.047 (0.014)	−0.048 (0.015)	−0.058 (0.011)	−0.059 (0.012)
SP ($\hat{\beta}_4$)	0.47 (0.24)	0.478 (0.223)	0.515 (0.181)	0.521 (0.19)

4. Discussion and Conclusions

4.1. Improving Traditional Regression Models

The usage of traditional regression models such as the GLMs may not always yield accurate results due to their assumptions of data independence, which does not hold true for trait data from related species. To overcome this, we have proposed a novel phylogenetic negative binomial regression model that takes into account the inherent dependence arising from shared ancestry. Estimations within this model are carried out using the GEE framework, ensuring a comprehensive analysis. Through rigorous bootstrapping simulations, we assess the model's effectiveness and demonstrate its practical application on empirical data obtained from a field study. Thus, our proposed model provides a more precise tool for analyzing count-dependent variables in related species, ultimately contributing to a more nuanced understanding of these relationships. Our model overcomes limitations inherent in traditional regression models, considering the inherent dependencies arising from shared ancestry. Hence, this model provides a robust and innovative tool for scientists conducting research on related species, thereby enhancing the quality and precision of findings in evolutionary biology and related fields.

4.2. Navigating Technical Challenges and Limitations

While our phylogenetic negative binomial regression model improves on traditional models such as GLM, it has limitations. Specifically, parameter estimation within our model requires advanced techniques to ensure rapid and convergent estimates. Moreover, our current model does not account for within-species variation, a scenario that warrants further exploration in future work. When assessing the robustness of our proposed phylogenetic negative binomial regression model, we estimated parameters using methods designed to solve systems of nonlinear equations. This was performed within a simulation under a bootstrapping approach. However, due to the high volume of repeated values in the simulated data, we occasionally encountered issues such as errors in singular value decomposition (SVD) when computing the inverse in the GEEs inverse in Equation (11). It is worth noting that these nonconvergent results reflect the inherent intricacies of the model and the potential limitations of the estimation method under specific conditions. In some cases, we had to reduce the tolerance during the decomposition of the equations to ensure a feasible estimate. However, this adjustment can occasionally lead to numerical instability, demonstrating the challenges inherent in balancing precision and computational stability in these complex models.

4.3. Exploring the Impact of Advanced Stochastic Processes in Phylogenetic Trait Evolution

The model can be expanded to accommodate additional evolutionary phenomena embedded in the covariate trait x . For instance, Paradis [9] assumed no elements from the correlation matrix C needed estimation, as they were directly derived from the given tree. However, expanding this assumption to include other processes with parameters within C

could provide a more comprehensive understanding of evolution. This could involve the Ornstein–Uhlenbeck process model with a force parameter α [50], Pagel’s λ model with a scale parameter λ [51], or the early burst model to allow for adaptive trait evolution [52].

The recent advancements in the field of stochastic processes have been substantial, particularly in the context of Ornstein–Uhlenbeck (OU) and fractional Brownian motion models. These models provide a comprehensive framework for analyzing trait evolution across phylogenetic trees, extending beyond the traditional scope of the Brownian motion stochastic process and accommodating greater complexity and flexibility [53,54]. Currently, models that incorporate random diffusivity have been explored [55–57]. In particular, the work of Wang et al. (2020) discusses residual non-ergodicity below the correlation time in the context of a fractional Brownian motion with random diffusivity [55]. Furthermore, they examined anomalous diffusion and non-ergodicity within heterogeneous diffusion processes using fractional Gaussian noise [58]. Meanwhile, models for scaled Brownian motion with random diffusivity were proposed by Dos Santos and Junior (2021) and Miyaguchi (2022) [56,57].

Building on these developments, the complex behaviors of these models have been the focus of more recent investigations. Hidalgo-Soria, Barkai, and Burov (2021) delved into the cusp of the non-Gaussian density of particles within a diffusing diffusivity model [59]. Additionally, Dos Santos, Menon Jr. and Cius (2022) utilized a superstatistical approach to explore the anomalous exponent for scaled Brownian motion [60]. These advances provide intriguing directions for future research. The influence of these more generalized stochastic processes on the results of our current study is an area of interest for further exploration. However, an in-depth examination of these models is beyond the scope of our current research. The implications of population size on the parameters of the underlying stochastic process, specifically the diffusion coefficient, are crucial, especially in large populations. Although these fluctuations might not drastically affect the results of empirical analyses, they could play a significant role in simulation studies. Our current study acknowledges the importance of these issues and advocates for their consideration in future research, further enriching the understanding of phylogenetic trait evolution.

4.4. Towards Multivariate Count Models

Further directions can be taken to build upon the present work. Initially, the focus can shift toward multivariate count models. While a univariate analysis offers a comprehensive understanding of single-variate distributions, multivariate analysis provides an understanding of interrelationships between multiple variables. In particular, it could be valuable to extend our efforts to multivariate Poisson or Negative Binomial distributions. A multivariate framework would permit the identification and examination of correlations between components. The *geepack* package’s `mmmmgee` function, as documented by [61], could facilitate this investigation.

4.5. Bayesian Approaches and Algorithmic Considerations

From a Bayesian viewpoint, there are also opportunities for further exploration. For example, we could expand on regression methods related to within-subject variability, as illustrated in the species examined by [14]. Another extension could involve adapting the multivariate Poisson log-normal model (PLN) proposed by [62]. This model connects p -dimensional observation vectors Y_j with Gaussian latent variables Z_j . Under PLN, these latent variables follow a normal distribution with a specified mean and covariance, while the observations adhere to an exponential distribution given these latent variables. The integration of phylogenetic tree structures within this model could allow for a more comprehensive analysis.

Looking ahead, we recognize the need to consider alternative algorithms that can better handle these challenging scenarios. In particular, the tree pruning algorithm [63] might offer an innovative way to alleviate these problems when applying the phylogenetic regression model to count dependent variables. Tree pruning could provide a way to

simplify the phylogenetic tree, thus reducing the computational burden and improving the stability of the model estimation process. This approach serves as an example of how future research could continue to refine these models, enhancing their robustness and accuracy in the analysis of data that exhibit dependencies due to shared ancestry.

5. Conclusions

In addressing the analytical challenges associated with dependent variables in related species, we have developed a phylogenetic negative binomial regression model that effectively addresses the analytical challenges associated with count-dependent variables in phylogenetically related species. This model, utilizing the GEE framework, not only rectifies the limitations of conventional GLMs but also enhances the quality and precision of findings in evolutionary biology. Importantly, this innovation opens up new avenues for future research, particularly in refining these models and exploring more advanced algorithms. These future directions are essential to better handle complex data dependencies and improve the robustness and accuracy of analyses.

Author Contributions: The authors' individual contributions are provided as the following: Conceptualization, D.-C.J.; methodology, C.-Y.W. and D.-C.J.; validation, D.-C.J.; formal analysis, C.-Y.W. and D.-C.J.; investigation, C.-Y.W. and D.-C.J.; resources, D.-C.J.; data curation, C.-Y.W. and D.-C.J.; writing—original draft preparation, D.-C.J.; writing—review and editing, D.-C.J.; visualization, C.-Y.W. and D.-C.J.; supervision, D.-C.J.; project administration, D.-C.J.; funding acquisition, D.-C.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research were funded by the National Science and Technology Council, Taiwan. NSTC 112-2118-M-035-003 -.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data can be accessed in Appendix A or by reaching out to the author.

Acknowledgments: We express our sincere gratitude to the editors and the four anonymous reviewers for their valuable suggestions, which have significantly improved the quality of this manuscript. We would like to express our gratitude to Brian O'Meara for his valuable suggestions and insights that greatly improved our work. Additionally, our appreciation goes to Hsun Jui (Ray) Chang for providing an optimal working environment during the revision process.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Scripts, Datasets, and Models

Appendix A.1. Code Availability

The script developed for this project can be accessed at the following link <https://tonyjhwueng.info/phy poinb2reg> (accessed on 18 August 2023).

- Table 1: https://tonyjhwueng.info/phy poinb2reg/mammal_lizard.html (accessed 27 July 2023).
- Figure 1: <https://tonyjhwueng.info/phy poinb2reg/illufigpropv2.html> (accessed on 27 July 2023).
- Figure 2: <https://tonyjhwueng.info/phy poinb2reg/schematicplot.html> (accessed on 27 July 2023).
- Table 2, Figure 3: <https://tonyjhwueng.info/phy poinb2reg/phy poisim.html> (accessed on 27 July 2023).
- Table 3, Figure 4: <https://tonyjhwueng.info/phy poinb2reg/phy nb2simb0b1box.html> (accessed on 27 July 2023).
- Figure 5: <https://tonyjhwueng.info/phy poinb2reg/lizardtree2.html> (accessed on 27 July 2023).
- Table 4: <https://tonyjhwueng.info/phy poinb2reg/lizardyoboot2.html> (accessed on 27 July 2023).

- Figure 6: <https://tonyjhwueng.info/phypoinb2reg/lizardyoboot3plot.html> (accessed on 27 July 2023).
- Figure 7, Table 5: <https://tonyjhwueng.info/phypoinb2reg/mammalyoboot2.html> (accessed on 27 July 2023).
- Table A1: <https://tonyjhwueng.info/phypoinb2reg/lizardyo2.html> (accessed on 27 July 2023).
- Table A2: <https://tonyjhwueng.info/phypoinb2reg/mammalplot.html> (accessed on 27 July 2023).

Appendix A.2. Trait Dataset

Appendix A.2.1. Lizard Trait Set

Table A1. Lizard trait set in [21]. EPY: eggs per year, EM: egg mass (g). This table depicts the means of life history count variables for all *Sceloporus* populations. Data sources, including life-history and mtDNA, are referenced in the final two columns [42].

No.	Species	EPY	EM
1	<i>S. undulatus</i> (GA)	23	0.33
2	<i>S. undulatus</i> (OH)	24	0.35
3	<i>S. undulatus</i> (AL)	25	0.28
4	<i>S. undulatus</i> (NJ)	18	0.36
5	<i>S. undulatus</i> (PA)	22	0.42
6	<i>S. undulatus</i> (SC)	22	0.33
7	<i>S. woodi</i>	12	0.25
8	<i>S. undulatus</i> (AZ)	25	0.29
9	<i>S. undulatus</i> (UT)	19	0.36
10	<i>S. undulatus</i> (Huerfano.CO)	22	0.32
11	<i>S. undulatus</i> (Mesa.CO)	16	0.42
12	<i>S. undulatus</i> (NE)	11	0.23
13	<i>S. undulatus</i> (TX)	28	0.22
14	<i>S. undulatus</i> (Grant.NM)	22	0.29
15	<i>S. undulatus</i> (Hidalgo.NM)	40	0.24
16	<i>S. virgatus</i>	10	0.23
17	<i>S. occidentalis</i>	11	0.50

Appendix A.2.2. Litter Size

Table A2. This table presents traits for 30 species [22], including litters number (LY), body size (LS), offspring value (OV), longevity (LG), and successful spread. Data are from [47,64], and other sources, encompassing variables from [65].

	Species	LY	LS	OV	Spread	LG
1	<i>Sciurus carolinensis</i>	2	2.98	0.02	1	24.00
2	<i>Sciurus vulgaris</i>	2	4.50	0.04	1	14.80
3	<i>Dipodomys ordii</i>	2	2.95	0.05	0	9.90
4	<i>Micromys minutus</i>	3	4.92	0.08	0	5.00
5	<i>Ondatra zibethicus</i>	2	6.55	0.05	1	10.00
6	<i>Mus musculus</i>	5	5.54	0.04	1	6.00
7	<i>Rattus exulans</i>	4	3.70	0.39	1	1.12
8	<i>Rattus norvegicus</i>	4	8.99	0.08	1	3.80
9	<i>Rattus rattus</i>	4	5.88	0.07	1	4.20
10	<i>Glis glis</i>	1	5.17	0.09	0	12.00
11	<i>Muscardinus avellanarius</i>	2	4.30	0.14	0	6.00
12	<i>Eliomys quercinus</i>	1	4.99	0.25	0	5.50
13	<i>Castor canadensis</i>	1	3.60	0.05	1	23.40

Table A2. Cont.

	Species	LY	LS	OV	Spread	LG
14	<i>Myodes glareolus</i>	4	4.31	0.06	1	4.92
15	<i>Myocastor coypus</i>	3	5.34	0.03	1	12.00
16	<i>Vulpes vulpes</i>	1	4.59	0.05	1	21.30
17	<i>Nyctereutes procyonoides</i>	1	6.36	0.07	1	16.60
18	<i>Procyon lotor</i>	1	3.06	0.05	1	21.00
19	<i>Mustela erminea</i>	1	6.74	0.08	1	12.50
20	<i>Mustela nivalis</i>	1	5.07	0.10	1	10.00
21	<i>Mustela putorius</i>	1	8.48	0.08	1	14.00
22	<i>Neovison vison</i>	1	4.50	0.09	1	11.40
23	<i>Genetta genetta</i>	2	2.29	0.02	1	34.00
24	<i>Herpestes javanicus</i>	2	2.21	0.05	0	10.00
25	<i>Equus caballus</i>	1	1.00	0.02	1	62.00
26	<i>Sus scrofa</i>	2	4.52	0.03	1	27.00
27	<i>Pecari tajacu</i>	2	1.56	0.02	0	31.50
28	<i>Camelus bactrianus</i>	1	1.39	0.03	1	40.00
29	<i>Hydropotes inermis</i>	1	3.00	0.08	0	13.90
30	<i>Muntiacus reevesi</i>	2	0.98	0.03	1	23.20

Appendix A.3. Count Regression for Independent Data

In Appendix A.3.1, we discuss the independent Poisson regression. The independent negative binomial regression is elucidated in Appendix A.3.2.

Appendix A.3.1. Independent Poisson Regression

In Poisson regression, we assume that the response variable y_i of the i th species is a count random variable following a Poisson distribution $\text{Poi}(\lambda)$, where λ_i is the expected frequency given a period of time [15]. The likelihood function given the comparative data $(y_1, y_2, \dots, y_n)$ under the independent case is

$$L(\lambda|y_1, y_2, \dots, y_n) = \prod_{i=1}^n \frac{\exp(-\lambda) \lambda^{y_i}}{y_i!}. \quad (\text{A1})$$

Let $\beta = (\beta_0, \beta_1, \dots, \beta_p)^t$, and X be the $n \times (p+1)$ design matrix composed by the column vectors $\mathbf{1}, x_1, \dots, x_p$ where $\mathbf{1} = (1, 1, \dots, 1)^t$ is a vector of 1 s and $x_j = [x_{ij}]$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$. The log-link function enables us to model the expected value of y_i as a linear combination of predictor variables. Consequently, the coefficient in Poisson regression represents the difference in the logarithm of expected counts per unit change in the corresponding predictor variable. By using the log link function on the mean, one has $\log(\lambda) = \log(E[y_i|x_{i0}, x_{i1}, x_{i2}, \dots, x_{ip}]) = \sum_{j=0}^p \beta_j x_{ij}$, and hence

$$\lambda = \exp\left(\sum_{j=0}^p \beta_j x_{ij}\right). \quad (\text{A2})$$

Taking the negative log of L in Equation (A1) and use Equation (A2), the negative log-likelihood function ℓ is shown in Equation (A3),

$$\begin{aligned} \ell(\beta|y, X) &= \sum_{i=0}^n (-\lambda + y_i \log(\lambda) - \log(y_i!)) \\ &= \sum_{i=0}^n \left(-\exp\left(\sum_{j=0}^p \beta_j x_{ij}\right) + y_i \sum_{j=0}^p \beta_j x_{ij} - \log(y_i!) \right). \end{aligned} \quad (\text{A3})$$

To obtain a maximum likelihood estimate for parameters $\hat{\beta}$, take the partial derivative of Equation (A3) with respect to β_j yields the Equation (A4)

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n x_{ij} \left(y_i - \exp \left(\sum_{j=0}^p \beta_j x_{ij} \right) \right), j = 0, 1, \dots, p. \quad (\text{A4})$$

Then, the MLE estimate of parameters $\hat{\beta}_i, i = 1, 2, \dots, p$ can be obtained by solving the system of equations in Equation (A4) through the numerical procedure such as Newton's method [17]. Readers can please refer to [66] for a more detailed description.

Appendix A.3.2. Independent Negative Binomial Regression

In negative binomial regression, we assume that the response variable y_i of the i th species is a random count variable following negative binomial distribution $\text{NB}(r, p)$ where $r > 0$ is the number of success until the experiment is stopped, y_i is the number of failures until the experiment is stopped and p is the probability of success in each experiment [67].

The likelihood function given the comparative data $\mathbf{y} = (y_1, y_2, \dots, y_n)$ under the independent case is

$$L(p, r | \mathbf{y}) = \prod_{i=1}^n \binom{y_i + r - 1}{y_i} p^r (1 - p)^{y_i}, \quad (\text{A5})$$

where $\binom{y_i + r - 1}{y_i} = \frac{\Gamma(r + y_i)}{\Gamma(y_i + 1) \Gamma(r)}$ and Γ is a gamma function [68,69].

Use the log link function $\log \frac{pr}{1-p} = \sum_{j=0}^p \beta_j x_{ij}$ for the mean, one has

$$p = \frac{1}{1 + r \exp(-\sum_{j=0}^p \beta_j x_{ij})}. \quad (\text{A6})$$

Let $\mu_i = \exp(\sum_{j=0}^p \beta_j x_{ij}) / (1 + \exp(\sum_{j=0}^p \beta_j x_{ij}))$. Taking the negative log of L in Equation (A5) and use Equation (A6), the negative log-likelihood function ℓ is shown in Equation (A7),

$$\begin{aligned} \ell(\beta | \mathbf{y}, \mathbf{X}) &= \sum_{i=1}^n \log \binom{y_i + r - 1}{y_i} p^r (1 - p)^{y_i} \\ &= \sum_{i=1}^n \log \binom{y_i + r - 1}{y_i} - (r + y_i) \log(1 + r\mu_i) + y_i \log(r\mu_i). \end{aligned} \quad (\text{A7})$$

To obtain the maximum likelihood estimate for parameters $\hat{\beta}$, taking the partial derivative to Equation (A7) with respect to β_j , namely, $\frac{\partial \ell}{\partial \beta_j}, j = 0, 1, 2, \dots, p$ as well as taking the derivative to Equation (A7) with respect to r , namely, $\frac{\partial \ell}{\partial r}$.

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^n \frac{x_{ij}(y_i - \mu_i)}{1 + \mu_i/r}, j = 0, 1, 2, \dots, p; \\ \frac{\partial \ell}{\partial r} &= \sum_{i=1}^n \left[r^2 \left(\log(1 + \frac{\mu_i}{r}) - \sum_{j=0}^{y_i-1} \frac{1}{j+r} \right) + \frac{r^2(y_i - \mu_i)}{r + \mu_i} \right]. \end{aligned} \quad (\text{A8})$$

Then, the MLE parameters $\hat{\beta}_j, j = 0, 1, \dots, p$ and $\hat{r}$ can be obtained by solving the system of equations in Equation (A8) through the numerical procedure such as Newton's method [17]. Readers can please refer to [67,69] for a more detail description.

Appendix A.4. Count Regression under GEE for Phylogenetic Dependent Data

The GEE is a strategy for parameter estimation in GLMs, accounting for potential correlation structures in outcomes, both known and unknown [15]. Traditional applications

of GEE propose several correlation structures, including independence, exchangeable, autoregressive order 1, and unstructured. The authors of [31] extended GEE to comparative count-dependent data analysis. For regression of phylogenetically dependent count data, we introduce a transformation matrix C derived from the phylogenetic tree to account for species interdependencies. Our primary models of interest are the Poisson regression and the negative binomial regression. Both models belong to the exponential family, where the associated random variable y has its density function [36].

$$f(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right). \quad (\text{A9})$$

Using GEE, we initial parameterize θ through a monotonic link function $g(\mu) = \theta$. We denote $E(y) = b'(\theta(\mu))$ as the mean function and $V(\mu) = b''(\theta(\mu))$ as the variance function which can be represented as differentials of $b(\theta)$, as indicated in Equation (A10)

$$\begin{aligned} E(y) &= b'(\theta) = \mu, \\ V(y) &= b''(\theta)a(\phi). \end{aligned} \quad (\text{A10})$$

Let $\eta_j = g(\mu_j) = X_j^t \beta$, given the interlinked relationships among the parameters, the partial derivative of ℓ with respect to β produces the estimation equation, guided by the chain rule expressed in Equation (A11).

$$\frac{\partial \ell}{\partial \beta} = \left[\left(\frac{\partial \ell}{\partial \theta} \right) \left(\frac{\partial \theta}{\partial \mu} \right) \left(\frac{\partial \mu}{\partial \eta} \right) \left(\frac{\partial \eta}{\partial \beta} \right) \right]_{(p+1) \times 1}. \quad (\text{A11})$$

The estimation equation for the regression parameter β is achieved by setting the derivative of the log-likelihood, ℓ , to zero which yields to the final estimating equation in Equation (A12).

$$\left[\left\{ \sum_{i=1}^n \left(\frac{y_i - \mu_i}{a(\phi)V(\mu_i)} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right)_i x_{ij} \right\}_{j=0,1,2,\dots,p} \right]_{(p+1) \times 1} = [\mathbf{0}]_{(p+1) \times 1}. \quad (\text{A12})$$

The following two subsections present our description to the GEE method for phylogenetic trait evolution [9]. We explore both the phylogenetic Poisson regression model [9] and the novel phylogenetic negative binomial regression model.

Appendix A.4.1 details the GEE application in the context of phylogenetic Poisson regression, while Appendix A.4.2 delves into its application for phylogenetic negative binomial regression. To our best knowledge, these regression models are not comprehensively covered in the existing literature. Our objective is to bridge this gap by providing an in-depth study of these regression models.

Appendix A.4.1. GEE for Phylogenetic Poisson Regression

In the following, we provide description for GEE procedure for the phylogenetic Poisson regression model (the reader can also refer to [9]).

Since the Poisson distribution with probability mass function $p(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$ belongs to the exponential family, refer to Equation (A9) one has

$$\begin{aligned} \theta &= \log \lambda, \\ a(\phi) &= 1, \\ b(\theta) &= \lambda, \\ c(y, \theta) &= -\log(y!). \end{aligned} \quad (\text{A13})$$

Then, the mean function and the variance function are shown in Equation (A14).

$$E(y) = V(y) = \lambda. \quad (\text{A14})$$

To apply the GEE for searching the regression estimate for β_j , since $\theta_i = \log \lambda_i$, $\lambda_i = \exp(\sum_{j=0}^p x_{ij}\beta_j) := \exp(\eta_i)$ and $\frac{\partial \lambda_i}{\partial \beta_j} = \sum_{j=0}^p x_{ij}\lambda_i$, $i = 1, 2, \dots, n$, the derivative of the negative log-likelihood function ℓ with respect to the parameter β_j can be computed by the chain rule in Equation (A12) which yields the partial derivative of the j th regression parameter in Equation (A15).

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \left[\frac{y_i - \lambda_i}{\lambda_i} \lambda_i x_{ij} \right], j = 0, 1, 2, \dots, p. \quad (\text{A15})$$

Let $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ and $\lambda = (\lambda_1, \dots, \lambda_n)$, one can deduce Equation (A15) in matrix form as

$$\left(\frac{\partial \lambda}{\partial \beta} \right)_{(p+1) \times n}^t V_{n \times n}^{-1} (y - \lambda)_{n \times 1}^t. \quad (\text{A16})$$

In the phylogenetic comparative regression model context, Ref. [9] defined the variance-covariance between the observation with

$$V = \phi A^{1/2} C A^{1/2} \quad (\text{A17})$$

where ϕ is dispersion parameter, C is the phylogenetic correlation matrix with elements c_{ij} of the shared branch length of taxa i and j and A is a diagonal matrix in Equation (A18).

$$A = \phi \cdot \text{diag}(V(y_i|x_i)) = \phi \cdot \text{diag}(\lambda) = \phi \cdot \text{diag}(\exp(\sum_{j=0}^p x_{ij}\beta_j)). \quad (\text{A18})$$

Therefore, given the response $Y = (y_1, y_2, \dots, y_n)^t$ and design matrix $X = (\mathbf{1}, x_1, x_2, \dots, x_p)^t$, by Equation (A12) the parameter estimate for β can be obtained by solving the nonlinear equation system of β .

$$X_{(p+1) \times n}^t \lambda_{n \times 1} \mathbf{1}_{1 \times n}^t (A_{n \times n}^{1/2} C_{n \times n} A_{n \times n}^{1/2})^{-1} (Y_{n \times 1} - \lambda_{n \times 1}) = [\mathbf{0}]_{(p+1) \times 1}. \quad (\text{A19})$$

Given trait data (x_{ij}, y_i) , $i = 1, \dots, n$; $j = 1, 2, \dots, p$ and the tree associated C matrix, an estimate of β_j s are estimated through a numerical search by solving the system of a nonlinear equation. The parameter estimate for β embedded can be obtained by solving the system of the nonlinear equation of β .

Appendix A.4.2. GEE for Negative Binomial Regression

For a negative binomial random variable y , $y \sim \text{NB}(r, p)$ where the parameter r is the number of successes and p is the probability of success in each trial. The probability mass function is $p(y|p, r) = \binom{y+r-1}{y} (1-p)^r p^y = \binom{y+r-1}{y} \exp(y \log p + r \log(1-p))$. The first moment is $E[y] = \mu = \frac{pr}{1-p}$ which shows that the p can be expressed in terms of μ and r as $p = \frac{\mu}{\mu+r}$.

Substitute p with $\mu/(\mu+r)$ into the probability mass function $p(y|p, r)$, we have

$$\begin{aligned} p(y|p, r) &= \binom{y+r-1}{y} \exp\left(y \log \frac{\mu}{\mu+r} + r \log \frac{r}{\mu+r}\right) \\ &= \exp\left(\frac{y \log \frac{\mu}{\mu+r} - r \log(\mu+r)}{1} + r \log r + \log \binom{y+r-1}{y}\right). \end{aligned} \quad (\text{A20})$$

Compare Equation (A20) with Equation (A9) one has $a(\phi) = 1$, $\theta = \log \frac{\mu}{\mu+r}$ which implies $\mu = r \exp(\theta) / (1 - \exp(\theta))$. Now consider $r \log(\mu + r) = r \log(r \exp(\theta) / (1 - \exp(\theta)) + r) = r \log(r / (1 - \exp(\theta))) = r \log r - r \log(1 - \exp(\theta))$, one has $b(\theta) = -r \log(1 - \exp(\theta))$, and $c(y, \phi) = 2r \log r + \log \left(\frac{y+r-1}{y} \right)$. Hence, we have

$$\begin{aligned}\theta &= \log \frac{\mu}{\mu+r}, \\ a(\phi) &= 1, \\ b(\theta) &= -r \log(1 - \exp(\theta)), \\ c(y, \phi) &= 2r \log r + \log \left(\frac{y+r-1}{y} \right).\end{aligned}\tag{A21}$$

The mean function $E[y]$ and the variance function can be written $E[y] = b'(\theta) = r \exp(\theta) / (1 - \exp(\theta))$, and the variance function $V[y]$ is $V[y] = b''(\theta)a(\phi) = r \exp(\theta) / (1 - \exp(\theta))^2$. That is,

$$\begin{aligned}E(y) &= b'(\theta) = \frac{r \exp(\theta)}{1 - \exp(\theta)} = \mu, \\ V(y) &= b''(\theta)a(\phi) = \frac{r \exp(\theta)}{(1 - \exp(\theta))^2} = \mu + \frac{\mu^2}{r}.\end{aligned}\tag{A22}$$

To apply the GEE [70] and numerical optimization for searching the regression estimate for β_i and r , since $r + \mu_i = r / (1 - \exp(\theta_i))$, one has $\frac{\mu_i}{r+\mu_i} = \exp(\theta_i)$. Then, consider to use log as the link function, one has $\theta_i = g(E[y_i]) = g(\mu_i) = \log \frac{\mu_i}{r+\mu_i}$ and variance function $V_{ii} = \mu_i + \mu_i^2 / r$, then $g(\mu_i) = \log \frac{\mu_i}{\mu_i+r} = \sum_{j=0}^p x_{ij}\beta_j$ implies

$$\begin{aligned}\mu_i &= r \frac{\exp(\sum_{j=0}^p x_{ij}\beta_j)}{(1 - \exp(\sum_{j=0}^p x_{ij}\beta_j))}, \\ V[E[y_i]] &= r \frac{\exp(\sum_{j=0}^p x_{ij}\beta_j)}{(1 - \exp(\sum_{j=0}^p x_{ij}\beta_j))^2}.\end{aligned}\tag{A23}$$

From Equation (A23), one has

$$\begin{aligned}\frac{\partial \mu_i}{\partial \beta_j} &= r x_{ij} \frac{\exp(\sum_{j=0}^p x_{ij}\beta_j)}{1 - \exp(\sum_{j=0}^p x_{ij}\beta_j)}, \\ \frac{\partial \mu_i}{\partial r} &= \frac{\exp(\sum_{j=0}^p x_{ij}\beta_j)}{1 - \exp(\sum_{j=0}^p x_{ij}\beta_j)}.\end{aligned}\tag{A24}$$

Let $\beta = (\beta_0, \beta_1, \dots, \beta_p)$, and one can deduce Equation (A23) in matrix form as in Equation (A25)

$$\begin{pmatrix} \frac{\partial \mu}{\partial \beta} \\ \frac{\partial \mu}{\partial r} \end{pmatrix} V^{-1} (y - X)\tag{A25}$$

In the phylogenetic comparative regression model context, Ref. [9] defined the variance-covariance between the observation with

$$V = \phi B^{1/2} C B^{1/2}\tag{A26}$$

where C is the phylogenetic correlation matrix with elements c_{ij} of the shared branch length of taxa i and j and B is a diagonal matrix with

$$B = \phi \cdot \text{diag}(V(y_i|x_i)) = \phi \cdot \text{diag}(\mu) = \phi \cdot \text{diag}\left(\frac{r \exp(\sum_{j=0}^p x_{ij}\beta_j)}{1 - \exp(\sum_{j=0}^p x_{ij}\beta_j)}\right). \quad (\text{A27})$$

Therefore, given the response $\mathbf{y} = (y_1, y_2, \dots, y_n)^t$ and the design matrix $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_p)^t$, by Equation (A12) the parameter estimate for β can be obtained by solving the nonlinear equation system of β . To incorporate C for deriving the phylogenetic negative binomial regression model, Equation (A12) can be expressed in a matrix form as shown in Equation (A28).

$$\mathbf{X}_{(p+1) \times n}^t \boldsymbol{\mu}_{n \times 1} \mathbf{1}_{1 \times n}^t (\mathbf{B}_{n \times n}^{1/2} \mathbf{C}_{n \times n} \mathbf{B}_{n \times n}^{1/2})^{-1} (\mathbf{Y}_{n \times 1} - \boldsymbol{\mu}_{n \times 1}) = [\mathbf{0}]_{(p+1) \times 1}. \quad (\text{A28})$$

The parameter estimate for β embedded in the μ vector with $\mu_i = r \exp(\sum_{j=0}^p x_{ij}\beta_j) / (1 - \exp(\sum_{j=0}^p x_{ij}\beta_j))$ can therefore be obtained through solving the system of the nonlinear equation of β .

References

- O'Meara, B.C. Evolutionary inferences from phylogenies: A review of methods. *Annu. Rev. Ecol. Evol. Syst.* **2012**, *43*, 267–285. [\[CrossRef\]](#)
- Adams, D.C. A generalized K statistic for estimating phylogenetic signal from shape and other high-dimensional multivariate data. *Syst. Biol.* **2014**, *63*, 685–697. [\[CrossRef\]](#)
- Freckleton, R.P.; Harvey, P.H.; Pagel, M. Phylogenetic analysis and comparative data: A test and review of evidence. *Am. Nat.* **2002**, *160*, 712–726. [\[CrossRef\]](#) [\[PubMed\]](#)
- O'Meara, B.; Heath, T.; Midford, P.E.; Chamberlain, S.; Brown, J.W.; Schliep, K. datelife: 0.2.3. *Datelife* **2016**, 56803. [\[CrossRef\]](#)
- Garamszegi, L.Z. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*; Springer: Berlin/Heidelberg, Germany, 2014.
- Harmon, L.J. *Phylogenetic Comparative Methods*; Independent: Traverse City, MI, USA, 2019.
- Davies, E.L.; Arbuckle, K. Coevolution of snake venom toxic activities and diet: Evidence that ecological generalism favours toxicological diversity. *Toxins* **2019**, *11*, 711. [\[CrossRef\]](#)
- Claramunt, S.; Hong, M.; Bravo, A. The effect of flight efficiency on gap-crossing ability in Amazonian forest birds. *Biotropica* **2022**, *54*, 860–868. [\[CrossRef\]](#)
- Paradis, E.; Claude, J. Analysis of comparative data using generalized estimating equations. *J. Theor. Biol.* **2002**, *218*, 175–185. [\[CrossRef\]](#)
- Lee, J.H.; Han, G.; Fulp, W.; Giuliano, A. Analysis of overdispersed count data: Application to the Human Papillomavirus Infection in Men (HIM) Study. *Epidemiol. Infect.* **2012**, *140*, 1087–1094. [\[CrossRef\]](#)
- Wu, C.Y. Phylogenetic Regression for Count Data: Negative Binomial Regression Model. Master's Thesis, Feng-Chia University, Taichung, Taiwan, 2022.
- Garre, A.; Zwietering, M.H.; van Boekel, M.A. The Most Probable Curve method-A robust approach to estimate kinetic models from low plate count data resulting in reduced uncertainty. *Int. J. Food Microbiol.* **2022**, *380*, 109871. [\[CrossRef\]](#)
- Mou, T.; Deng, W.; Gu, F.; Pawitan, Y.; Vu, T.N. Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing. *Front. Genet.* **2020**, *10*, 1331. [\[CrossRef\]](#)
- Contreras-Reyes, J.E.; Quintero, F.O.L.; Wiff, R. Bayesian modeling of individual growth variability using back-calculation: Application to pink cusk-eel (*Genypterus blacodes*) off Chile. *Ecol. Model.* **2018**, *385*, 145–153. [\[CrossRef\]](#)
- Agresti, A. *An Introduction to Categorical Data Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
- Cleophas, T.J.; Zwinderman, A.H. Poisson regression for binary outcomes (52 Patients). In *SPSS for Starters and 2nd Levelers*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 273–277.
- Burden, R.L.; Faires, J.D. *Numerical Analysis*, 8th ed.; Thomson Brooks/Cole: Pacific Grove, CA, USA, 2005.
- Hafemeister, C.; Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **2019**, *20*, 296. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chen, C.W.; Khamthong, K. Bayesian modelling of nonlinear negative binomial integer-valued GARCHX models. *Stat. Model.* **2020**, *20*, 537–561. [\[CrossRef\]](#)
- Brooks, M.E.; Kristensen, K.; Darrigo, M.R.; Rubim, P.; Uriarte, M.; Bruna, E.; Bolker, B.M. Statistical modeling of patterns in annual reproductive rates. *Ecology* **2019**, *100*, e02706. [\[CrossRef\]](#)
- Niewiarowski, P.H.; Angilletta, M.J.; Leaché, A.D. Phylogenetic comparative analysis of life-history variation among populations of the lizard *Sceloporus undulatus*: An example and prognosis. *Evolution* **2004**, *58*, 619–633. [\[CrossRef\]](#) [\[PubMed\]](#)

22. Hassler, G.; Tolkoﬀ, M.R.; Allen, W.L.; Ho, L.S.T.; Lemey, P.; Suchard, M.A. Inferring phenotypic trait evolution on large trees with many incomplete measurements. *J. Am. Stat. Assoc.* **2020**, *17*, 678–692. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Burnham, K.P.; Anderson, D.R.; Huyvaert, K.P. AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behav. Ecol. Sociobiol.* **2011**, *65*, 23–35. [\[CrossRef\]](#)
24. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: New York, NY, USA, 2002.
25. Cavanaugh, J.E. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Stat. Probab. Lett.* **1997**, *33*, 201–208. [\[CrossRef\]](#)
26. Jhwueng, D.C. On the covariance of phylogenetic quantitative trait evolution models and their matrix condition. *Commun.-Stat.-Simul. Comput.* **2022**, 1–20. [\[CrossRef\]](#)
27. Felsenstein, J. Phylogeny and the comparative method. *Am. Nat.* **1985**, *125*, 1. [\[CrossRef\]](#)
28. Stamatakis, A. A review of approaches for optimizing phylogenetic likelihood calculations. In *Bioinformatics and Phylogenetics: Seminal Contributions of Bernard Moret*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 1–19.
29. Revell, L.J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **2012**, *3*, 217–223. [\[CrossRef\]](#)
30. Pennell, M.; Eastman, J.; Slater, G.; Brown, J.; Uyeda, J.; Fitzjohn, R.; Alfaro, M.; Harmon, L. geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* **2014**, *30*, 2216–2218. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Paradis, E.; Claude, J.; Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **2004**, *20*, 289–290. [\[CrossRef\]](#) [\[PubMed\]](#)
32. CRAN Task View: Phylogenetics, Especially Comparative Methods. Available online: <https://cran.r-project.org/web/views/Phylogenetics.html> (accessed on 27 July 2023).
33. Ané, C. Analysis of comparative data with hierarchical autocorrelation. *Evolution* **2008**, *2*, 1078–1102. [\[CrossRef\]](#)
34. Jhwueng, D.C. Assessing the Goodness of Fit of Phylogenetic Comparative Methods: A Meta-Analysis and Simulation Study. *PLoS ONE* **2013**, *8*, e67001. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Pfennig, D.W. *Phenotypic Plasticity & Evolution: Causes, Consequences, Controversies*; Taylor & Francis: Milton Park, UK, 2021.
36. Hardin, J.W.; Hilbe, J.M. *Generalized Estimating Equations*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2002.
37. Liang, K.Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22. [\[CrossRef\]](#)
38. Efron, B. Resampling plans and the estimation of prediction error. *Stats* **2021**, *4*, 1091–1115. [\[CrossRef\]](#)
39. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2002; ISBN 0-387-95457-0.
40. O’Meara, B.; Ané, C.; Sanderson, M.; Wainwright, P. Testing different rates of continuous trait evolution using likelihood. *Evolution* **2006**, *60*, 922–933.
41. Jhwueng, D.C.; O’Meara, B.C. On the Matrix Condition of Phylogenetic Tree. *Evol. Bioinform.* **2020**, *16*, 1176934320901721. [\[CrossRef\]](#)
42. Leaché, A.D.; Reeder, T.W. Molecular systematics of the eastern fence lizard (*Sceloporus undulatus*): A comparison of parsimony, likelihood, and Bayesian approaches. *Syst. Biol.* **2002**, *51*, 44–68. [\[CrossRef\]](#)
43. Sanderson, M.J. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **2003**, *19*, 301–302. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Yu, G. *Data Integration, Manipulation and Visualization of Phylogenetic Trees*; CRC Press: Boca Raton, FL, USA, 2022.
45. Míguez, S.; Torre, I.; Arrizabalaga, A.; Freixas, L. Influences of Maternal Weight and Geographic Factors on Offspring Traits of the Edible Dormouse in the NE of the Iberian Peninsula. *Life* **2023**, *13*, 1223. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Sherman, P.W.; Braude, S.; Jarvis, J.U. Litter sizes and mammary numbers of naked mole-rats: Breaking the one-half rule. *J. Mammal.* **1999**, *80*, 720–733. [\[CrossRef\]](#)
47. Capellini, I.; Baker, J.; Allen, W.L.; Street, S.E.; Venditti, C. The role of life history traits in mammalian invasion success. *Ecol. Lett.* **2015**, *18*, 1099–1107. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Stoltzfus, A.; Lapp, H.; Matasci, N.; Deus, H.; Sidlauskas, B.; Zmasek, C.M.; Vaidya, G.; Pontelli, E.; Cranston, K.; Vos, R.; et al. Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. *BMC Bioinform.* **2013**, *14*, 158. [\[CrossRef\]](#)
49. Fritz, S.A.; Bininda-Emonds, O.R.; Purvis, A. Geographical variation in predictors of mammalian extinction risk: Big is bad, but only in the tropics. *Ecol. Lett.* **2009**, *12*, 538–549. [\[CrossRef\]](#)
50. Hansen, T.F. Stabilizing selection and the comparative analysis of adaptation. *Evolution* **1997**, *51*, 1341–1351. [\[CrossRef\]](#)
51. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **1999**, *401*, 877. [\[CrossRef\]](#)
52. Harmon, L.J.; Losos, J.B.; Jonathan Davies, T.; Gillespie, R.G.; Gittleman, J.L.; Bryan Jennings, W.; Kozak, K.H.; McPeck, M.A.; Moreno-Roark, F.; Near, T.J.; et al. Early bursts of body size and shape evolution are rare in comparative data. *Evolution* **2010**, *64*, 2385–2396. [\[CrossRef\]](#)
53. Uhlenbeck, G.E.; Ornstein, L.S. On the theory of the Brownian motion. *Phys. Rev.* **1930**, *36*, 823. [\[CrossRef\]](#)
54. Deng, W.; Barkai, E. Ergodic properties of fractional Brownian-Langevin motion. *Phys. Rev. E* **2009**, *79*, 011112. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Wang, W.; Cherstvy, A.G.; Chechkin, A.V.; Thapa, S.; Seno, F.; Liu, X.; Metzler, R. Fractional Brownian motion with random diffusivity: Emerging residual nonergodicity below the correlation time. *J. Phys. A Math. Theor.* **2020**, *53*, 474001. [\[CrossRef\]](#)

56. dos Santos, M.A.; Junior, L.M. Random diffusivity models for scaled Brownian motion. *Chaos Solitons Fractals* **2021**, *144*, 110634. [CrossRef]
57. Miyaguchi, T. Generalized Langevin equation with fluctuating diffusivity. *Phys. Rev. Res.* **2022**, *4*, 043062. [CrossRef]
58. Wang, W.; Cherstvy, A.G.; Liu, X.; Metzler, R. Anomalous diffusion and nonergodicity for heterogeneous diffusion processes with fractional Gaussian noise. *Phys. Rev. E* **2020**, *102*, 012146. [CrossRef]
59. Hidalgo-Soria, M.; Barkai, E.; Burov, S. Cusp of non-Gaussian density of particles for a diffusing diffusivity model. *Entropy* **2021**, *23*, 231. [CrossRef]
60. dos Santos, M.; Menon, L., Jr.; Cius, D. Superstatistical approach of the anomalous exponent for scaled Brownian motion. *Chaos Solitons Fractals* **2022**, *164*, 112740. [CrossRef]
61. Ristl, R.; McDaniel, L.; Henderson, N.; Prague, M. *Mmmgee: Simultaneous Inference for Multiple Linear Contrasts in GEE Models*; R CRAN: Vienna, Austria, 2018.
62. Aitchison, J.; Ho, C. The multivariate Poisson-log normal distribution. *Biometrika* **1989**, *76*, 643–653. [CrossRef]
63. Posada, D.; Crandall, K.A. Felsenstein phylogenetic likelihood. *J. Mol. Evol.* **2021**, *89*, 134–145. [CrossRef]
64. Ernest, S.M. Life history characteristics of placental nonvolant mammals: Ecological archives E084-093. *Ecology* **2003**, *84*, 3402. [CrossRef]
65. Bielby, J.; Mace, G.M.; Bininda-Emonds, O.R.; Cardillo, M.; Gittleman, J.L.; Jones, K.E.; Orme, C.D.L.; Purvis, A. The fast-slow continuum in mammalian life history: An empirical reevaluation. *Am. Nat.* **2007**, *169*, 748–757. [CrossRef] [PubMed]
66. Stanford University. STATS 200: Introduction to Statistical Inference Autumn 2016 Lecture 27—Poisson Regression. 2016. Available online: <https://web.stanford.edu/class/archive/stats/stats200/stats200.1172/Lecture27.pdf> (accessed on 22 July 2023).
67. NCSS Statistical Software. Chapter 326: Negative Binomial Regression. 2023. Available online: https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative_Binomial_Regression.pdf (accessed on 22 July 2023).
68. Wikipedia contributors. Gamma function—Wikipedia, The Free Encyclopedia. 2023. Available online: https://en.wikipedia.org/wiki/Gamma_function (accessed on 22 July 2023).
69. Agresti, A. *Foundations of Linear and Generalized Linear Models*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
70. Kong, M.; Xu, S.; Levy, S.M.; Datta, S. GEE type inference for clustered zero-inflated negative binomial regression with application to dental caries. *Comput. Stat. Data Anal.* **2015**, *85*, 54–66. [CrossRef] [PubMed]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.