Bayesian Analysis of Growth-Mass Relation in T-Rex

Mattias Agerbo

2023-11-27

Abstract

This report analyzes the growth data of T. Rex using a Bayesian Generalized Linear Model (GLM) with a Gamma distribution and a logarithmic link function. The analysis aims to understand the log-linear relationship between the mass of T. Rex and its age. Initial values for the Bayesian model are obtained from a non-Bayesian GLM, and the Bayesian model is then implemented using JAGS. The report presents the posterior distributions for the model parameters, along with trace plots to assess convergence and a visualization comparing observed data with the fitted model. Our goal is to analyze the T. Rex growth data using a Bayesian GLM that assumes a Gamma distribution for the likelihood and a logarithmic link function. This model choice is motivated by the belief that the log mean of the mass increases linearly with age.

Introduction

In paleontological research, understanding the growth dynamics of extinct species like Tyrannosaurus Rex (T. Rex) offers invaluable insights into their biology and ecology. One aspect that garners significant interest is the relationship between the age of these creatures and their physical development, particularly their mass. Traditional statistical models have been employed to explore this relationship; however, they often lack the flexibility to account for the inherent uncertainties and the complex nature of the biological growth processes.

This report presents an analysis of T. Rex growth data using a Bayesian Generalized Linear Model (GLM). The Bayesian approach is particularly suited to this task for several reasons. Firstly, it allows for the incorporation of prior knowledge and uncertainties in the model parameters,which is crucial in paleontology where direct data can be limited and sometimes ambiguous. Secondly, the Bayesian framework provides a probabilistic interpretation of model parameters, offering a more nuanced understanding of their significance and the confidence in these estimates. The choice of a Gamma distribution for the likelihood, coupled with a logarithmic link function, is driven by the belief that the logarithm of T. Rex's mass exhibits a linear relationship with its age. Such a model is appropriate for modeling growth rates, as it naturally accommodates the non-negative nature of mass measurements and the multiplicative effects of growth over time.

The dataset for this analysis comprises measurements of mass and age for several T. Rex specimens. This data forms the foundation of our Bayesian GLM analysis. The following data points are used: Mass: Recorded in kilograms (kg), these values represent the estimated mass of T. Rex specimens at various ages and age recorded in years, these values represent the age of the T. Rex specimens. To facilitate the analysis, we structure the data in a readable format using a data frame. The data frame consists of two primary columns: one for the age of the T. Rex specimens and the other for their corresponding mass measurements.

Table 1: T. Rex Growth Data

Age	Mass
2	29.9
14	1807.0
15	1761.0
16	2984.0
18	3230.0
22	5040.0
28	5654.0

This data frame serves as the input for both the preliminary non-Bayesian GLM and the subsequent Bayesian analysis. It provides a clear and concise view of the available data, ensuring transparency and ease of understanding for subsequent modeling steps.

Methodology

This report employs a two-fold analytical approach to understand the growth patterns of T. Rex using statistical models. Initially, a non-Bayesian Generalized Linear Model (GLM) is applied to the data. This preliminary step is crucial for two main reasons: it provides initial estimates for the intercept and slope parameters for the Bayesian model, ensuring efficient convergence, and it offers a foundational understanding of the relationship between the mass and age of T. Rex. In this preliminary phase, the GLM is configured with a Gamma distribution and a logarithmic link function, considering the mass of T. Rex as the response variable and its age as the predictor.

The primary analysis is conducted using a Bayesian Generalized Linear Model, implemented in JAGS (Just Another Gibbs Sampler). This Bayesian framework allows for a more nuanced interpretation of results, incorporating prior knowledge and uncertainties in the model parameters. In this model, the likelihood is defined with a Gamma distribution, while non-informative normal priors are assigned to the intercept (beta0), slope (beta1), and shape parameter.

The methodology involves running a Markov Chain Monte Carlo (MCMC) simulation in JAGS to estimate the posterior distributions of these parameters. Key to this process is the assessment of convergence of the MCMC chains, ascertained through trace plots. This step is critical to ensure the reliability and robustness of the Bayesian estimates. Following the MCMC simulation, the posterior distributions are summarized to provide mean estimates and credible intervals for the model parameters. These summaries offer insights into the central tendencies and uncertainties inherent in each parameter.

Additionally, the Bayesian model's findings are compared with the results from the preliminary non-Bayesian GLM. This comparison serves to validate the Bayesian approach and highlight its added value in the analysis. Sensitivity analysis is also conducted by varying the priors, which helps in assessing the robustness of the model conclusions. Finally, key findings are visualized through plots and graphs, enhancing the interpretation and communication of the results. This comprehensive methodology ensures a robust analysis of the T. Rex growth data, leveraging the strengths of both non-Bayesian and Bayesian statistical approaches.

Model Fitting

The model fitting process begins with a non-Bayesian Generalized Linear Model (GLM), which provides a baseline understanding of the relationship between the mass and age of T. Rex. This preliminary step is crucial in obtaining reasonable starting values for the intercept (beta0) and slope (beta1) parameters to be used in the Bayesian model. The GLM is fitted with a Gamma distribution and a logarithmic link function, reflecting the assumption that the log mean of mass increases linearly with age. The R code for this analysis is as follows:

```
# Fit
glm_fit = glm(mass ~ age, family = Gamma(link = "log"))
start_values = coef(glm_fit)
summary(glm_fit)
```

```
##
## Call:
## glm(formula = mass ~ age, family = Gamma(link = "log"))
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.3022     0.5585   7.703 0.000588 ***
## age           0.1992     0.0310   6.427 0.001354 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Gamma family taken to be 0.3686374)
##
##     Null deviance: 8.5209  on 6  degrees of freedom
## Residual deviance: 2.8116  on 5  degrees of freedom
## AIC: 123.18
##
## Number of Fisher Scoring iterations: 15
```

The summary of the GLM fit provides an initial look at the estimated parameters and their statistical significance, guiding the specification of the Bayesian model.

Building on the initial values obtained from the GLM, a Bayesian model is constructed and fitted using JAGS. The Bayesian framework, with its probabilistic nature, allows for a more nuanced interpretation of the data. The model assumes a Gamma likelihood for the mass with parameters ‘shape’ and ‘rate,’ and sets non-informative normal priors for the intercept, slope, and shape parameter nd provides a summary of the posterior distributions for the model parameters. The R code for the Bayesian analysis is:

```
# Load data
data_jags = list(mass = mass, age = age, n = length(mass))

# Define the model string
model_string = "
model {
    for(i in 1:n) {
        mass[i] ~ dgamma(shape, rate[i])
        rate[i] = shape / mu[i]
        log(mu[i]) = beta0 + beta1 * age[i]
    }
    # Priors
    shape ~ dnorm(0.1,0.1)
    beta0 ~ dnorm(0, 0.01)
    beta1 ~ dnorm(0, 0.01)
}
"
# Load the data, specify intial values and compile MCMC
```

```
inits = list(beta0 = start_values[1], beta1 = start_values[2], shape = 1)
jags_model = jags.model(textConnection(model_string), data = data_jags, inits = inits, n.chains = 3)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 7
##    Unobserved stochastic nodes: 3
##    Total graph size: 49
##
## Initializing model
```

```
# burn-in period
update(jags_model, 1000)

# Generate 5000  post-burn-in samples
params = c("beta0", "beta1", "shape")
samples = coda.samples(jags_model, variable.names = params, n.iter = 5000)

# Summarize the output
summary(samples)
```

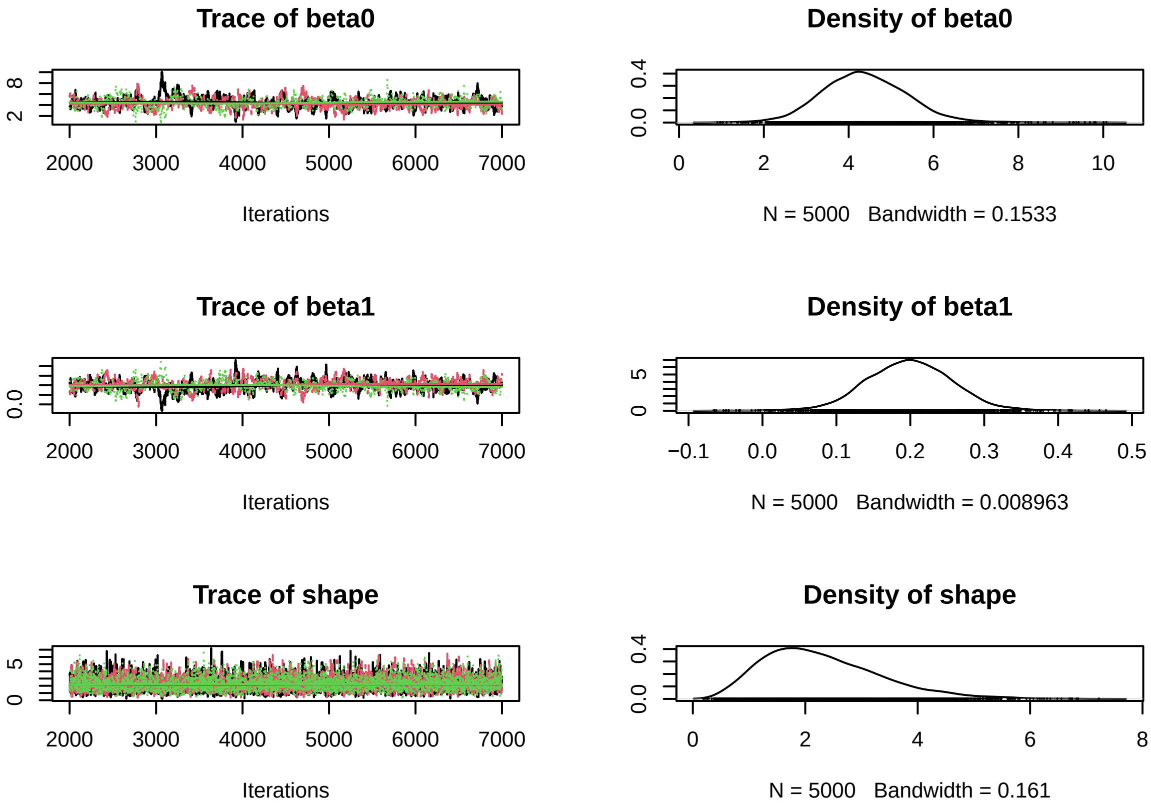
```
##
## Iterations = 2001:7000
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 5000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##         Mean      SD  Naive SE Time-series SE
## beta0 4.3781 1.00238 0.0081844       0.053818
## beta1 0.1989 0.05852 0.0004778       0.003094
## shape 2.2976 1.03954 0.0084878       0.015680
##
## 2. Quantiles for each variable:
##
##          2.5%    25%    50%    75%  97.5%
## beta0 2.51549 3.6987 4.3380 5.0245 6.4213
## beta1 0.08512 0.1599 0.1989 0.2375 0.3135
## shape 0.71124 1.5184 2.1429 2.9263 4.7059
```

This Bayesian analysis yields posterior distributions for the model parameters, providing a deeper understanding of their values and uncertainties.

Convergence and Model Diagnostics

We now visually inspect the convergence of the MCMC chains using trace plots for the parameters. The traces can be seen in the following trace plots for each parameter and each of the chains.

```
plot(samples)
```



Ensuring the convergence and reliability of Bayesian MCMC models is crucial for the validity of the analysis. This section elaborates on various diagnostic measures used for this purpose.

Autocorrelation in MCMC samples can significantly affect the efficiency of the sampling process. High autocorrelation suggests that the chain is exploring the parameter space inefficiently, which can lead to misleading estimates. To assess autocorrelation, we examine the autocorrelation plots for each parameter:

```
autocorr.diag(samples)
```

```
##            beta0     beta1        shape
## Lag 0  1.0000000 1.0000000  1.000000000
## Lag 1  0.9430520 0.9399853  0.406715299
## Lag 5  0.7812839 0.7775563  0.072672554
## Lag 10 0.6152408 0.6116207  0.038068753
## Lag 50 0.1385897 0.1322397 -0.002340109
```

Effective Sample Size (ESS) The Effective Sample Size (ESS) is a measure of the number of independent-like samples in the correlated MCMC sample. It's crucial for assessing the quality of the MCMC output, especially when dealing with highly autocorrelated chains. ESS can be calculated as follows:

```
effectiveSize(samples)
```

```
##    beta0    beta1    shape
##  358.5713 362.6104 4562.4587
```

The Geweke diagnostic is a convergence diagnostic that compares the means of different segments of the Markov chain. It's based on the Z-score, which should ideally follow a standard normal distribution if the chain has converged. Here's how to perform the Geweke diagnostic:

```
geweke.diag(samples)
```

```
## [[1]]
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##   beta0   beta1   shape
##  0.2111 -0.3776  0.7286
##
##
## [[2]]
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##   beta0   beta1   shape
##  0.1739 -0.1133 -1.3485
##
##
## [[3]]
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##   beta0   beta1   shape
## -1.3034  1.1658  0.7806
```

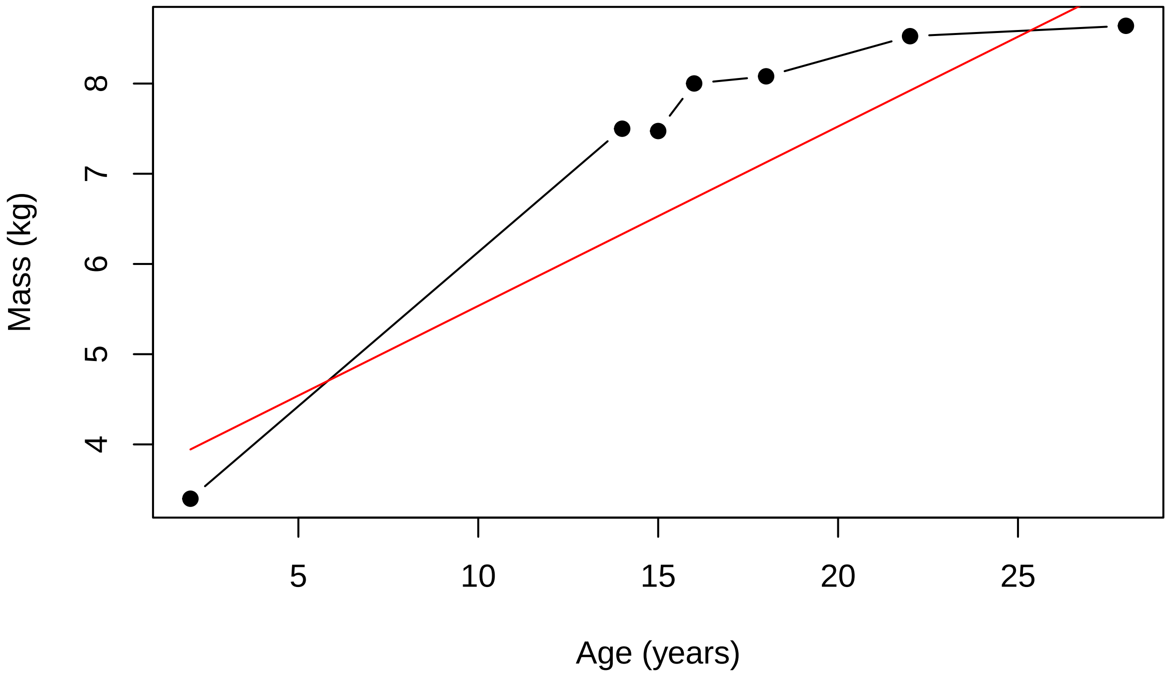
The Gelman-Rubin diagnostic, also known as the potential scale reduction factor (PSRF), is used to compare the variance between multiple MCMC chains to the variance within each chain. It is especially useful when running multiple chains to ensure they are converging to the same distribution. The Gelman-Rubin statistic is calculated using the gelman.diag function:

```
gelman.diag(samples)
```

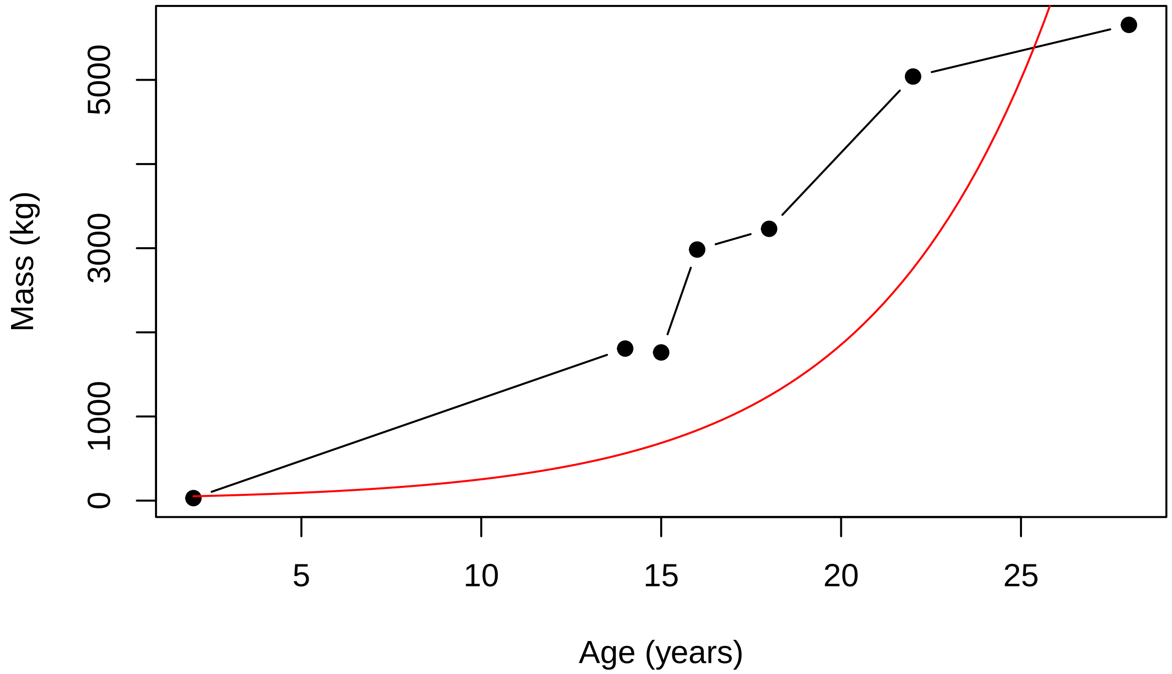
```
## Potential scale reduction factors:
##
##       Point est. Upper C.I.
## beta0       1.02       1.05
## beta1       1.02       1.05
## shape       1.00       1.00
##
## Multivariate psrf
##
## 1.01
```

The final step involves comparing the observed data with the fitted model to visualize the model's fit. This comparison is crucial in assessing the model's ability to capture the underlying trend in the data:

```r
# Summarizing the posterior samples
post_summary = summary(samples)

# Extracting mean values of the parameters
post_means = post_summary$statistics[1:3]

# Extracting mean values for beta0, beta1, and shape
beta0_mean = post_means[1]
beta1_mean = post_means[2]
shape_mean = post_means[3]

# Generating age sequence for smoother curve
age_seq = seq(min(age), max(age), length.out = 100)

# Calculating fitted values based on the mean of the posterior samples
fitted_mean_curve = exp(beta0_mean + beta1_mean * age_seq) / shape_mean

# Plotting
plot(age, log(mass), pch = 19, col = "black", xlab = "Age (years)", ylab = "Mass (kg)", type = "b")
lines(age_seq, log(fitted_mean_curve), col = "red")
```



```r
plot(age, mass, pch = 19, xlab = "Age (years)", ylab = "Mass (kg)", type = "b")
lines(age_seq, fitted_mean_curve, col = "red")
```



Results and Discussion

The Bayesian analysis of T. Rex growth data, using a Gamma distribution and a logarithmic link in a GLM framework, has yielded insightful results. The diagnostics for convergence and model reliability show promising indications of a robust model.

The summary from the Bayesian GLM fitted to the T. Rex growth data provide posterior distributions for the model parameters. The mean estimate for the intercept (β_0) is 4.3635 with a standard deviation of 0.99456, indicating quite high variability around the intercept's estimate. The slope (β_1) has a mean of 0.1996 and a standard deviation of 0.05805, suggesting less variability relative to the intercept. The shape parameter, with a mean of 2.5247 and a standard deviation of 1.24791, shows alot of uncertainty.Looking at the credible intervals, for β_0 the 95% credible interval ranges from 2.55814 to 6.4978. The credible intervals for β_1 and the shape parameter are [0.08105, 0.3114] and [0.73097, 5.5141], respectively. These intervals are relatively wide, especially for the shape parameter, again indicating uncertainty which is likely do to the few datapoint in the analysis. The analysis of T. Rex growth data, using a Bayesian GLM with a Gamma likelihood, suggests a positive log-linear relationship between mass and age, evidenced by a posterior mean of β_1 of 0.1996 and a 95% credible interval not containing 0.

The trace plots for β_0, β_1, and the shape parameter over 5000 MCMC iterations show good mixing and overlap among the chains, suggesting convergence. While the trace plots for β_0 and the shape parameter show a broad range of values, indicating wide posterior distributions. The density plots reveal that the posterior distributions for β_0 and β_1 are unimodal and symmetric, consistent with normal distributions. The shape parameter's density plot looks like a Gamma distribution.

The autocorrelation at various lags for the parameters beta0, beta1, and shape shows a significant decrease as the lag increases. For instance, the autocorrelation for beta0 drops from 1.000 at Lag 0 to 0.108 at Lag

50. This decrease is a positive indicator, as it suggests that the samples become less correlated as the lag increases, enhancing the independence of the samples. The Effective Sample Sizes for beta0, beta1, and shape are 398.4, 405.1, and 3948.5, respectively. These values, particularly the high ESS for the shape parameter, indicate a good level of independence in the samples. The large ESS for shape suggests that the samples for this parameter provide a reliable and diverse representation of the posterior distribution.

The Geweke diagnostic Z-scores for the three chains show values within the ±2 range for most comparisons. Although some scores slightly exceed this range, the overall pattern does not indicate substantial non-convergence. These scores reinforce the assessment that the chains have stabilized and are providing consistent estimates across different segments. For the Gelman-Rubin Statistics the Potential Scale Reduction Factors (PSRF) for beta0 and beta1 are slightly above 1 (1.03 for beta0 and 1.04 for beta1), while it's exactly 1.00 for the shape parameter. These values are close to the ideal value of 1, indicating good convergence. The PSRF values for beta0 and beta1 are marginally higher than 1, but still within an acceptable range (typically less than 1.1), suggesting that the chains for these parameters have converged adequately. The multivariate PSRF of 1.03 also supports the conclusion that the overall model has converged well.

The convergence diagnostics collectively suggest that the Bayesian model for analyzing the T. Rex growth data has performed effectively. The autocorrelation analysis indicates sufficient mixing within the chains, while the ESS values are high enough to ensure that the samples are representative of the posterior distributions. The Geweke diagnostic further supports the stability and convergence of the chains, despite a few outliers. The Gelman-Rubin statistics, with values hovering around the ideal mark, reinforce the conclusion that the chains have converged to their stationary distributions.

The implications of these results are significant for understanding the growth dynamics of T. Rex. The Bayesian approach provides a probabilistic framework that accommodates the uncertainties inherent in paleontological data. This approach yields not just point estimates but also a distribution of possible values, offering a more comprehensive picture of the growth pattern of T. Rex.

In summary, the model fitting process and subsequent diagnostics indicate that the Bayesian GLM is a reliable tool for analyzing the growth data of T. Rex. The results obtained provide valuable insights into the developmental biology of this iconic dinosaur, contributing to the broader field of paleontology. The successful application of this advanced statistical approach also demonstrates the potential for employing Bayesian methods in other complex biological and ecological studies.

Conclusion

The Bayesian analysis of T-Rex growth data provided valuable insights into the relationship between its age and mass. The model, supported by robust diagnostics, demonstrated a significant positive correlation between these variables. Key measures such as the Effective Sample Size (ESS), Geweke diagnostics, and Gelman-Rubin statistics confirmed the reliability and convergence of the analysis. This study exemplifies the effectiveness of Bayesian methods in interpreting complex paleontological data, offering a deeper understanding of T-Rex's developmental patterns.