A Simple Growth Problem Using Bayesian Estimates

1 Introduction

A growth curve is an empirical graphical representation of the evolution of a quantity over time. Growth curves are widely used in statistics to determine the type of growth pattern of the quantity—be it linear, exponential, Weibull, etc. Once the type of growth is determined, mathematical model can be created to make inferences on the model assumed parameters and predict future observations.

Multilevel modeling and structural equation modeling are the two primary analysis techniques for growth curve analysis. Multilevel model is also known as hierarchical linear models, mixed models, and other terms. Hierarchical regression models are based on the disaggregation of the model into multiple levels of explanation and thus, are important once there are predictors at different levels of variation or in the analysis of data are obtained by stratified or cluster sampling [1].

Our project focused on the growth curve for rats. Only one predictor is considered in our models: time. Several model can be proposed, including non-hierarchical normal model and hierarchical normal models. For hierarchical models, both uncorrelated and correlated normal models could be reasonable. The bivariate normal hierarchical model assumes correlated birth-weight and growth-rate for each rat that come from a bivariate normal prior whereas the other does not. From a biological standpoint, the correlation between birth-weight and growth-rate might exist as a result of the genetic effects. Therefore, our study investigated the bivariate cases and try to provide a provision of the potential genetic effects. Two scenarios were discussed here: without and with parameter extension models. Gibbs samplers were derived. However, with the extended parameters model, we do encounter the singular matrix problem which hinder the Gibbs samplings for certain parameters. Metropolis–Hastings (MH) algorithm with sliced Gibbs Sampling were also attempted, we did not reach to a reasonable solution using R. Hence, Openbugs was employed here as it also adapts Hamiltonian Monte Carlo algorithm (HMC) to fulfill the simulations. The difference between SSE from observed y and SSE from replicated y based on the models were simulated for model checking. Deviance information criteria (DIC) was chosen to compare the model fitting of our two models. The reasons for the insignificant linear correlation between the birth-weight and growth-rate is discussed.

2 Statistical Models and Simulation Approaches

Our data was taken from Gelfand and Hills (1990) table 3 [2]. 30 young rats were weighted weekly for 5 weeks and time was recorded in days. Figure 1 indicates a potential increasing trend for the weight over time. Thus, the linear regression model is chosen as the modeling tool. The regression model in a matrix form is given as

$$\begin{aligned} y_{ij} &= X_j\theta_i + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma^2) \\ y_{ij} &\sim N(X_j\theta_i, \sigma^2) \end{aligned}$$

where $X_{j_{1\times 2}} = (1\ x_j)$, $\theta_i = (\alpha_i, \beta_i)$, for $i = 1, ..., 30,\ j = 1, ..., 5$. Here, α_i and β_i are the birth-weight and growth-rate for each rat.

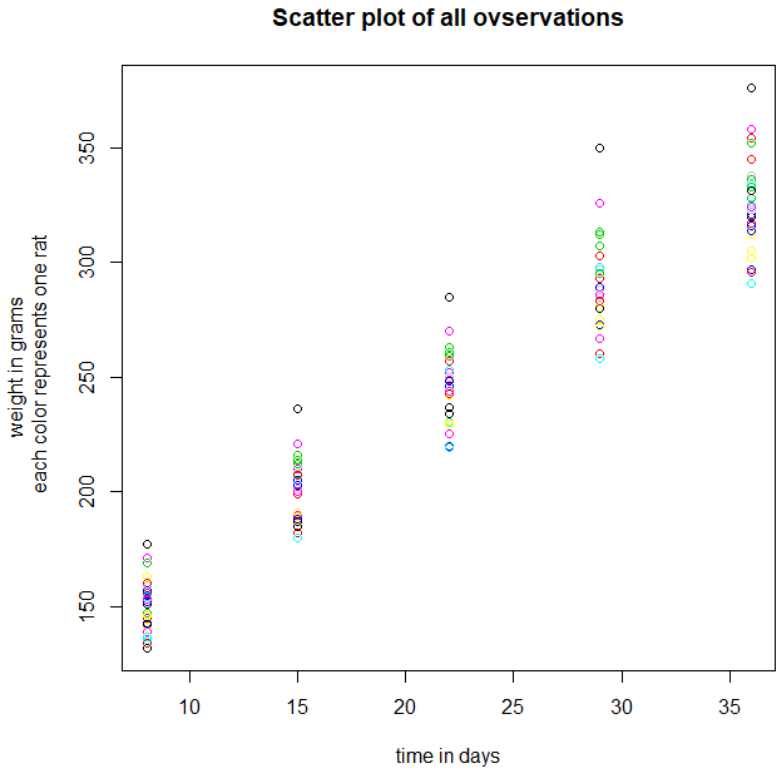

Figure 1: Scatter plot for all observations.

Model 1

Let's consider the regression coefficients are directly from a multivariate distribution, i.e.

$$\theta_i \sim MVN(\mu_\theta = \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \Sigma_\theta = \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}).$$

Non-informative prior for μ_θ and σ^2 is chosen. In addition, a conjugate hyperprior distribution for Σ_θ is chosen as

$$\Sigma_\theta \sim Inv - Wishart_3(I_{2\times 2}).$$

The correlation between the birth-weight and growth-rate is derived as

$$\rho = \frac{\Sigma_{\theta_{12}}}{\sqrt{\Sigma_{\theta_{11}}\Sigma_{\theta_{22}}}}.$$

Then the Gibbs samplers can be derived as follows.

$$p(\theta_i|\theta_{-i},\mu_\theta,\Sigma_\theta,\sigma^2,y) \sim MVN((\sum_{j=1}^{5} X_j^T X_j \sigma^{-2} + \Sigma_\theta^{-1})^{-1}(\sum_{j=1}^{5} X_j^T y_{ij}\sigma^{-2} + \Sigma_\theta^{-1}\mu_\theta),\ \text{precision} = \sum_{j=1}^{5} X_j^T X_j \sigma^{-2} + \Sigma_\theta^{-1})$$

$$\text{for } i = 1, ..., 30.$$

$$p(\mu_\theta|\theta,\Sigma_\theta,\sigma^2,y) \sim MVN(\bar{\theta}, \frac{\Sigma_\theta}{30})$$

$$p(\Sigma_\theta|\theta,\mu_\theta,\sigma^2,y) \sim Inv - Wishart_{33}((S+I)^{-1}) \quad \text{where } S = \sum_{i=1}^{30}(\theta_i - \mu_\theta)(\theta_i - \mu_\theta)^T$$

$$p(\sigma^2|\theta,\mu_\theta,\Sigma_\theta,y) \sim Inv - Gam(74, \frac{1}{2}\sum_{i,j}(y_{ij} - X_j\theta_i)^2))$$

Model 2

The trouble with the above Inv-Wishart$_3(I)$ model is that it strongly constrains the variance parameters, the diagonal elements of the covariance matrix [1]. A model that is noninformative on the correlations but allows a wider range of uncertainty on the variances can be established as,

$$y_{ij} \sim N(X_j\theta_i, \sigma^2), \quad \text{for } i = 1, ..., 30,\ j = 1, ..., 5$$
$$\text{split } \theta_i = \mu_\theta + \xi \otimes \eta_i,\ \text{where } \otimes \text{ is element-wise multiplication.}$$

Now a conjugate hyperprior distribution for Σ_η is chosen as $\Sigma_\eta \sim Inv - Wishart_3(I_{2\times 2})$. The correlation between the birth-weight and growth-rate is derived as

$$\rho = \frac{\xi_1\xi_2\Sigma_{\eta_{12}}}{|\xi_1\xi_2|\sqrt{\Sigma_{\eta_{11}}\Sigma_{\eta_{22}}}}.$$

The covariance matrix for the regression coefficients is $\Sigma_\theta = diag(\xi)\Sigma_\eta diag(\xi)$. Then, we have

$$p(\eta_i|\eta_{-i},\mu_\theta,\xi,\Sigma_\eta,\sigma^2,y) \propto exp(-\frac{1}{2\sigma^2}\sum_{j=1}^{5}(X_j(\xi\otimes\eta_i) - (y_{ij} - X_j\mu_\theta))^T(X_j(\xi\otimes\eta_i) - (y_{ij} - X_j\mu_\theta)))$$

$$p(\mu_\theta|\eta,\xi,\Sigma_\eta,\sigma^2,y) \propto exp(-\frac{1}{2\sigma^2}\sum_{i,j}(X_j\mu_\theta - (y_{ij} - X_j(\xi\otimes\eta_i)))^T(X_j\mu_\theta - (y_{ij} - X_j(\xi\otimes\eta_i))))$$

$$p(\xi|\eta,\mu_\theta,\Sigma_\eta,\sigma^2,y) \propto exp(-\frac{1}{2\sigma^2}\sum_{i,j}(X_j(\xi\otimes\eta_i) - (y_{ij} - X_j\mu_\theta))^T(X_j(\xi\otimes\eta_i) - (y_{ij} - X_j\mu_\theta)))$$

$$p(\Sigma_\eta|\eta,\mu_\theta,\xi,\sigma^2,y) \sim Inv - Wishart_{33}((S+I)^{-1}), \quad \text{where } S = \sum_{i=1}^{30}\eta_i\eta_i^T$$

$$p(\sigma^2|\theta,\mu_\theta,\Sigma_\eta,y) \sim Inv - Gam(74, \frac{1}{2}\sum_{i,j}(y_{ij} - X_j(\mu_\theta + \xi\otimes\eta_i))^2).$$

Gibbs samplers for η_i, μ_θ and ξ cannot be achieved as the inverse of $X_j^T X_j$ does not exist. MH simulation were also attempted, but we do not achieve a reasonable solution. Thus, Openbugs is involved here.

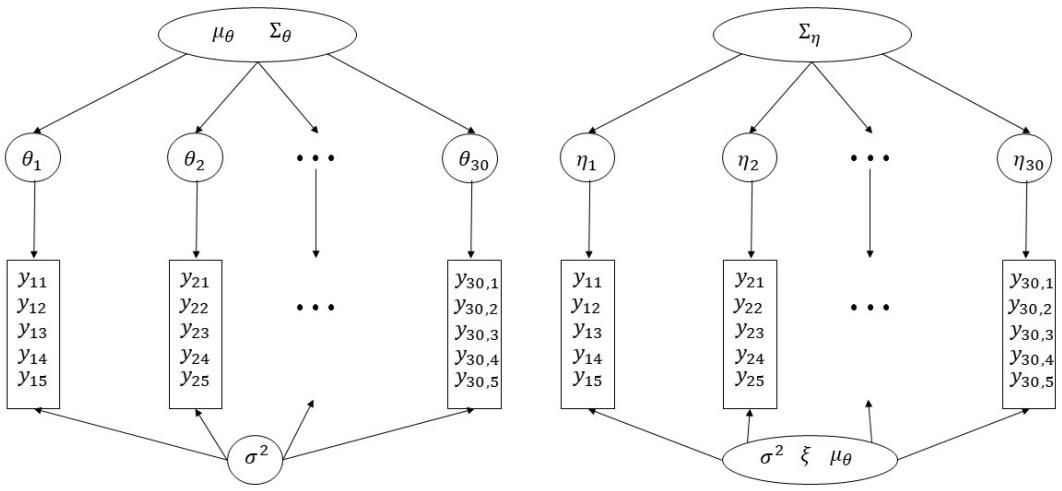

Figure 2: Hierarchical models for the rats growth curve studies. Left: Model 1. Right Model 2. It is obviously that the parameter space for the model 2 is larger due to extra parameter ξ.

3 Bayesian Inferences

Model 1

The Bayesian inferences for ρ, μ_θ, σ, and Σ_θ (which is given by its inverse R) is given by table 1 along with DIC. 1000 iterations were performed in which 500 iterations were used as warmups. The potential scale reduction factor on rank normalized split chains (Rhat) for all parameters are < 1.05 indicating convergence. Also, the posterior density plots have been examined to be smooth and consistent with the conditional posterior distributions derived in previous section.

Table 1: Bayesian Inference for the input samples of Model 1

	Q5	Q50	Q95	Mean	SD	Rhat	Bulk_ESS	Tail_ESS
rho	-0.473	-0.105	0.403	-0.086	0.258	1.009	181	207
mu.theta[1]	102.9	106.6	110.2	106.53	2.314	1.003	596	601
mu.theta[2]	6.021	6.187	6.364	6.187	0.103	1	742	807
R[1,1]	0.005	0.009	0.018	0.01	0.004	1.009	165	286
R[1,2]	-0.118	0.019	0.092	0.01	0.063	1.01	193	180
R[2,1]	-0.118	0.019	0.092	0.01	0.063	1.01	193	180
R[2,2]	2.473	4.139	7.81	4.54	1.822	1.008	272	272
sigma	5.407	6.115	6.943	6.135	0.478	1.002	246	522

	Dbar	Dhat	DIC	pD
Y	986.2	928.1	1044	58.16

Model 2

The Bayesian inferences for ρ, μ_θ, σ, ξ and Σ_η (which is given by its inverse R) is given by table 2 along with DIC.

Table 2: Bayesian Inference for the input samples of Model 2

	Q5	Q50	Q95	Mean	SD	Rhat	Bulk_ESS	Tail_ESS
rho	-0.457	-0.108	0.29	-0.099	0.228	1.006	1280	2339
xi[1]	0.567	1.21	2.25	1.283	0.502	1.03	45	96
xi[2]	0.929	1.839	3.115	1.906	0.667	1.101	22	61
mu.theta[1]	101.995	106.3	110.3	106.279	2.568	1.022	114	471
mu.theta[2]	6.007	6.184	6.358	6.183	0.107	1.034	101	420
R[1,1]	0.003	0.013	0.039	0.016	0.011	1.025	53	145
R[1,2]	-0.135	0.035	0.23	0.041	0.119	1.011	558	462
R[2,1]	-0.135	0.035	0.23	0.041	0.119	1.011	558	462
R[2,2]	3.515	13.13	35.763	15.508	10.377	1.074	29	73
sigma	5.422	6.081	6.921	6.121	0.461	1.001	1821	2855

	Dbar	Dhat	DIC	pD
Y	968.2	941.5	994.8	26.67

5000 iterations were performed in which 2500 iterations were used as warmups. At convergence, the potential scale reduction factor on rank normalized split chains (Rhat) for each parameter should be ≤ 1.05. However,

$$Rhat_{\xi_2} = 1.101,$$
$$Rhat_{R_{22}} = 1.074,$$

indicating, there is potential convergence problem for our extended parameter model. Two reasons may account for this: either the model is not a good fit for the observations or there are not enough observations for the large parameter space.

4 Discussions

4.1 Results

At the beginning of model checking stage, we replicated a large number of y values for each model.

Suppose the model is a good fit, then we would expect the difference between SSE from observed y and SSE from replicated y will follow a normal distribution with mean around zero.

Figure 3 shows that under model 1, the difference for the SSE does roughly follow a normal distribution. Under model 2, the the difference for the SSE is also approximately normally distributed around mean zero (Figure 4). No discrepancy has been observed. Model 2 is expected to be a better fit for the rats data since the shape of its distribution is more symmetric.

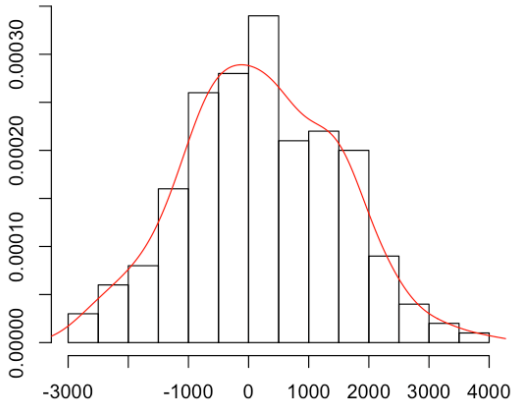

Figure 3: Histogram of the difference of SSE between the data and Model 1 replicates.

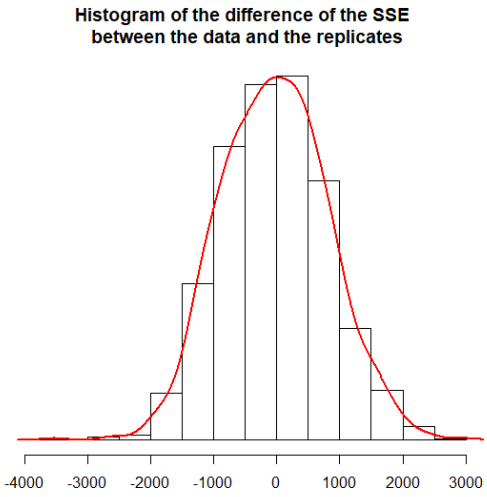

Figure 4: Histogram of the difference of SSE between the data and Model 2 replicates.

From table 3, we see the DIC for model 2 is smaller than that for model 1, suggesting model 2 to be a better fit for the rats data. This is consistent with the observation on the SSE differences. The extended parameter does provide more flexibility for modeling the variances of the regression.

Table 3: DIC comparison for both models

	Dbar	Dhat	**DIC**	pD
Model1	977.1	922.4,	**1031.8**	54.70
Model2	968.2	941.5	**994.8**	26.67

In addition, the bivariate normal hierarchical regression model assumed a homogeneous linear correlation ρ between the regression coefficients. Whereas, the Bayesian inference provides an insignificant

linear correlation ρ (Table 1, 2). Therefore, the assumptions for this particular model may not be very appropriate for our observed data.

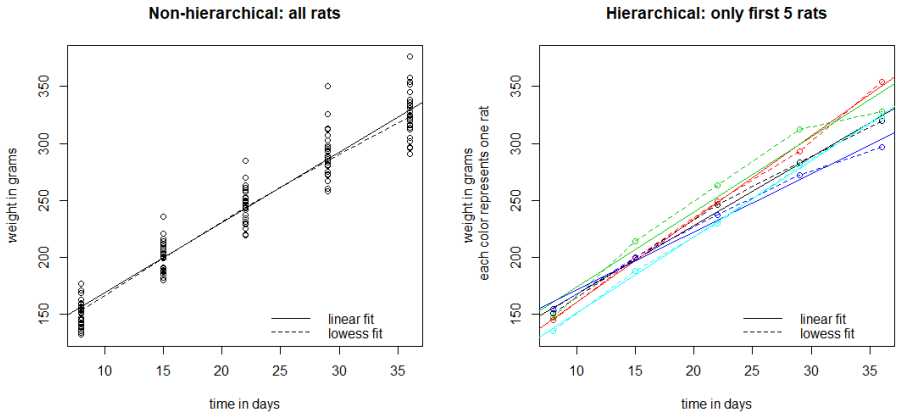

Figure 5: Compassion for linear regression model and non-linear model. The dataset does show slight curvature trend regardless of non-hierarchical or hierarchical model.

4.2 Future work

Further steps are required to investigate the growth curve. Although an obvious increasing trend for the weight over time exists, it does differ slightly from the linear trend (Figure 5). From biological standpoint, the correlation between birth-weight and growth-rate is of primary interest. Thus, we might model the weight using hierarchical logistic regression growth curve as follows.

$$y_{ij} = \frac{\alpha_i}{1+\beta_i e^{-\gamma_i x_j}} + \epsilon_{ij},\ \epsilon_{ij} \sim N(0,\sigma^2),\ \text{for } i = 1,...,30;\ j = 1,...,5.$$

Apparently,

$$\begin{aligned}
\text{birth weight}_i &= \frac{\alpha_i}{1+\beta_i}, \\
\text{growth rate}_i &= \frac{\alpha_i\beta_i\gamma_i e^{-\gamma_i x_j}}{(1+\beta_i e^{-\gamma_i x_j})^2}, \\
\text{life-weight-limit}_i &= \alpha_i \text{ when time} \to \infty.
\end{aligned}$$

Now, the growth-rate is a function of birth-weight and time, yielding an evolving correlation between birth-weight and growth-rate. Weakly informative priors can be chosen as

$$\alpha_i \sim N(\mu_\alpha, \sigma_\alpha^2),\ \beta_i \sim N(\mu_\beta, \sigma_\beta^2),\ \gamma_i \sim N(\mu_\gamma, \sigma_\gamma^2),$$

where σ_α^2, σ_β^2, and σ_γ^2 are large. Similar approaches can be performed by using the Gompretz growth curve, Richard's growth curve and etc with weakly informative normal priors (i.e. large variances).

Gompretz hierarchical model:

$$y_{ij} \sim N(\alpha_i e^{-\beta_i e^{-\gamma_i x_j}}, \sigma^2), \text{ for } i = 1, ..., 30; \ j = 1, ..., 5.$$
$$\text{birth weight}_i = \alpha_i e^{-\beta_i},$$
$$\text{growth rate}_i = \alpha_i \beta_i \gamma_i e^{-\beta_i e^{-\gamma_i x_j} - \gamma_i x_j},$$
$$\text{life-weight}_{i \ time \to \infty} = \alpha_i.$$

Richard's hierarchical model:

$$y_{ij} \sim N(\frac{\alpha_i}{(1 + \beta_i e^{-\gamma_i x_j})^{\frac{1}{\delta_i}}}, \sigma^2), \text{ for } i = 1, ..., 30; \ j = 1, ..., 5.$$
$$\text{birth weight}_i = \frac{\alpha_i}{(1 + \beta_i)^{\frac{1}{\delta_i}}},$$
$$\text{growth rate}_i = \frac{\alpha_i \beta_i \gamma_i e^{-\gamma_i x_j}}{\delta_i (1 + \beta_i e^{-\gamma_i x_j})^{\frac{1}{\delta_i} + 1}},$$
$$\text{life-weight}_{i \ time \to \infty} = \alpha_i.$$

Apparently, no simple Gibbs Samplers can be attained. Hence, HMC can be employed to make inferences. All three models mentioned above involve exponential terms in the model. The increasing trends in a young age range between the linear model and other models does not differ a lot, however the exponential term in non-linear models do provide the time-flexibility for the correlation, which is believed to be the truth by biologists. However, all three models mentioned above suffer serious exploitation problem as we only have data for young rats. Therefore potential convergence or poor fit issues are expected. Under convergence, the difference for SSE between the observed data and the replicated data under the model can be simulated to check the model, and DIC can be used to chose the model if discrepancy does not occur. Also, the parameter spaces for these models are relatively large comparing with the 150 sample size.

Although the hierarchical models discussed here may contribute to more variations for the data, non-hierarchical model for certain parameters can be considered if the Bayesian inferences provide quite similar results among groups. This will decrease the parameter space and increase the power.

5 Conclusion

In this project, we have analyzed a simple growth curve problem under the hierarchical linear model setting using Bayesian approach. We also assumed bi-variate normal distribution as priors for the parameter vector in both models. For the first model, we did it the regular way, i.e. there is no parameter extension for θ. However, we split θ as $\mu_\theta + \xi \otimes \eta$ for the second model for the purpose of allowing more uncertainty on variances of θ. The first model performed well overall, while the second one had a bit convergence problem. The extended parameter space can explain more variations for the rats data. Furthermore, all estimates for ρ are around 0 indicating there is no linear relationship between birth weight and growth rate in our sample data.

As a conclusion, our bi-variate normal assumption with linear relationship between intercept and slope parameters turns out to be inappropriate. In fact, if we were able to obtain more sample data in a longer period of time, an approximate sigmoid or "S"-shaped curve could be observed which is more reasonable in a biological perspective. Therefore, we may need to further try some non-linear hierarchical models in order to reach an appropriate result.

References

[1] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[2] Alan E. Gelfand, Susan E. Hills, Amy Racine-Poon, and Adrian F. M. Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, 85(412):972–985, 1990.