Journal of Data Science **2**(2004), 259-272

Using the Box-Cox Power Transformation to Predict Temporally Correlated Longitudinal Data

R. C. Hwang
Dahan Institute of Technology

Abstract: In this paper, the repeated measurement linear model proposed by Diggle (1988) is applied to two real data examples to predict future values for temporally correlated longitudinal data. This model incorporates the population mean, variability among individuals, serial correlation within an individual, and measurement error. In practice, however, the original data may not fit well with the linearity assumption imposed on the mean function by Diggle's model, thereby deteriorating the overall prediction ability of the model. To overcome this potential drawback, the Box-Cox power transformation (Box and Cox 1964) is considered, and two different ways of conducting power transformations are suggested. One of these two approaches performs transformation inside of Diggle's model, and the other performs transformation outside of Diggle's model. Given Diggle's model using the power transformed data, two prediction methods (the maximum likelihood method and the approximate Bayesian approach) are used to predict future values. Using our real data examples, it is shown that both values of mean absolute difference and mean absolute relative difference for each of these two prediction methods without power transformation can be reduced by more than 10% by simply performing power transformation. Results indicate that the prediction ability of Diggle's model can be significantly improved by employing power transformation, because lower levels of both mean absolute difference and mean absolute relative difference can be obtained.

Key words: Approximate Bayesian approach, Box-Cox power transformation, inverse gamma distribution, maximum likelihood method, noninformative prior, repeated measurement linear model.

1. Introduction

The repeated measurement linear model proposed by Diggle (1988) is an extremely popular method of predicting future values for temporally correlated longitudinal data. The chief advantage of this model is that it incorporates simultaneously the population mean, variability among individuals, serial correlation

within an individual, and measurement error. In this paper, Diggle's model is applied to two real data examples. When tailor-made to fit our data structure as described in Section 3, the model can be defined by

$$Y_j = X\beta + \tau_j \mathbf{1} + V_j + \epsilon_j, \tag{1.1}$$

for each $j = 1, \cdots, n$. Here, Y_j is a $p \times 1$ random vector representing p observations made at equally spaced times $(1, \cdots, p)$ on the jth subject,

$$X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & p \end{pmatrix}^T$$

a known $p \times 2$ design matrix, and $\beta = (\beta_1, \beta_2)^T$ an unknown 2×1 vector of regression coefficients. The τ_j is a normal random variable with mean 0 and variance σ_τ^2, $\mathbf{1}$ a $p \times 1$ vector of 1's, V_j a $p \times 1$ normal random vector with mean 0 and covariance matrix $\sigma_V^2 C$, and ϵ_j a $p \times 1$ normal random vector with mean 0 and covariance matrix $\sigma_\epsilon^2 I$. Here, $0 < \sigma_\tau^2, \sigma_V^2, \sigma_\epsilon^2 < \infty$, C stands for a $p \times p$ correlation matrix, and I is a $p \times p$ identity matrix. Finally, for $j = 1, \cdots, n$, τ_j, V_j and ϵ_j are all independent. For other related prediction methods, see, for example, the monograph by Diggle, Liang and Zeger (1994).

By (1.1), through a straightforward calculation, the covariance matrix of Y_j can be expressed by

$$Cov(Y_j) = \sigma_V^2 \ \Sigma,$$

where $\Sigma = \phi_1 \ \mathbf{1} \ \mathbf{1}^T + C + \phi_2 \ I$, $\phi_1 = \sigma_\tau^2/\sigma_V^2$ and $\phi_2 = \sigma_\epsilon^2/\sigma_V^2$. To predict future values, the covariance structure in V_j generally has to be extensible to the future values of the individuals observed. In this paper, the dependence structure in V_j is taken to be the autoregressive process of order 1 (Lee 1988); hence,

$$C = \left(\phi_3^{|i-j|}\right), \tag{1.2}$$

for $i, j = 1, \cdots, p$, where $\phi_3 \in (-1, 1)$ is the AR(1) parameter.

Using the model specified by (1.1) and (1.2), the purpose of this paper is to predict the future value y_j at the design point $x = (1, p+1)$, given the total observed data $Y = (Y_1^T, \cdots, Y_n^T)^T$. Here, y_j is a random variable representing a future observation to be made at time $p+1$ on the jth subject, for each $j = 1, \cdots, n$. This is a time series prediction, and is therefore important in practice. Under these circumstances, the estimation of parameters and the prediction of future values have been considered in Diggle (1988) and Donnelly, Laird and Ware (1995), respectively, by using the maximum likelihood method (MLM).

Note that the mean function $X\beta$ in the right-hand side of (1.1) is linear. In practice, however, the original data may not fit well with such linearity assumptions. In such a case, the prediction ability of the associated model may deteriorate. To avoid this possible situation, we apply the Box-Cox power transformation

(Box and Cox 1964) to the original data. When the value of the Box-Cox power transformation parameter is chosen adequately, the resulting transformed data can fit better with the linearity assumption imposed on the mean function by Diggle's model, thus making Diggle's model more appropriate for the transformed data.

The Box-Cox power transformation method is described accordingly. Set the real number λ as the Box-Cox power transformation parameter. Given the value of λ, the power transformed value $Y_j^{(\lambda)}$ for Y_j is a $p \times 1$ vector and its kth element $Y_{jk}^{(\lambda)}$ is defined by

$$\begin{aligned} Y_{jk}^{(\lambda)} &= \{(Y_{jk} + \omega_{jk})^{\lambda} - 1\}/\lambda, && \text{for } \lambda = 0. \\ &= \log(Y_{jk} + \omega_{jk}), && \text{for } \lambda \neq 0, \end{aligned}$$

Here, Y_{jk} is the kth element of Y_j, and ω_{jk} is a known constant such that $Y_{jk} + \omega_{jk} > 0$, for each $k = 1, \cdots, p$. In practice, $\omega_{jk} = 0$, if Y_{jk} is positive.

In practice, two different ways of choosing the value of λ are suggested to perform power transformation. The first approach treats λ as a parameter of Diggle's model; then, following the idea of (1.1), it fits $Y_j^{(\lambda)}$ as

$$Y_j^{(\lambda)} = X\beta + \tau_j \mathbf{1} + V_j + \epsilon_j, \tag{1.3}$$

for each $j = 1, \cdots, n$. The second approach does not treat λ as a parameter of Diggle's model. It simply and directly transforms the original data such that the power transformed data set have the best linear fit. Specifically, it takes the minimizer λ^* of the function TSSE(λ) defined below as the selected value of λ. Given the value of λ, the total sum of squared error TSSE(λ) (Neter, Wasserman and Kutner 1989) is defined by

$$TSSE(\lambda) = \sum_{j=1}^{n} \|Y_j^{(\lambda)} - X\hat{\beta}_j(\lambda)\|^2.$$

Here, $\hat{\beta}_j(\lambda) = \arg\ \min_{\beta} \|Y_j^{(\lambda)} - X\beta\|^2$, X and β have been given in (1.1), and the notation $\|\cdot\|$ stands for the Euclidean norm. When the value of λ^* is obtained, we fit the power transformed data $Y_j^{(\lambda^*)}$ for (1.1) accordingly

$$Y_j^{(\lambda^*)} = X\beta + \tau_j \mathbf{1} + V_j + \epsilon_j, \tag{1.4}$$

for each $j = 1, \cdots, n$.

This article is organized as follows. Section 2 employs both the MLM and the approximate Bayesian approach (ABA) to predict y_j, given each model specified

by (1.2) and (1.3), and by (1.2) and (1.4). Section 3 illustrates the prediction ability of Diggle's model using two real data examples.

2. Prediction

In this section, both the MLM and the ABA are employed to predict the future value y_j, given the observed data Y. Their formulations for the model specified by (1.2) and (1.3) are given in Subsections 2.1 and 2.2, respectively; and those for the model described by (1.2) and (1.4) are contained in Subsection 2.3. For the sake of simplicity, assume that Y_{jk} are positive, for all j and k.

2.1 The MLM

According to the results obtained by applying the MLM to both (1.2) and (1.3), the likelihood function of β, σ_V^2, ϕ_1, ϕ_2, ϕ_3 and λ, given Y, can be expressed by

$$L(\beta, \sigma_V^2, \phi_1, \phi_2, \phi_3, \lambda | Y) \propto (\det\ \sigma_V^2 \Sigma)^{-n/2} \times$$
$$\exp[-1/(2\sigma_V^2) \sum_{j=1}^{n} \{Y_j^{(\lambda)} - X\beta\}^T \Sigma^{-1} \{Y_j^{(\lambda)} - X\beta\}]\ |J|. \tag{2.1}$$

Here, $J = \prod_{j=1}^{n} \prod_{k=1}^{p} Y_{jk}^{\lambda-1}$, the notation detA denotes the determinant of the matrix A, and $a_n \propto b_n$ means that there is a normalizing constant c such that $a_n = c\ b_n$.

By (2.1) and the idea of generalized least squares (Seber 1977), given ϕ_1, ϕ_2, ϕ_3 and λ, the maximum likelihood estimators of β and σ_V^2 can be expressed respectively as

$$\begin{aligned} \tilde{\beta}(\phi_1, \phi_2, \phi_3, \lambda) &= (n\ X^T \Sigma^{-1} X)^{-1} \{\sum_{j=1}^{n} X^T \Sigma^{-1} Y_j^{(\lambda)}\}, \\ \tilde{\sigma}_V^2(\phi_1, \phi_2, \phi_3, \lambda) &= (np)^{-1} \sum_{j=1}^{n} \{Y_j^{(\lambda)} - X\tilde{\beta}\}^T \Sigma^{-1} \{Y_j^{(\lambda)} - X\tilde{\beta}\}. \end{aligned}$$

Fitting these results into (2.1), through a straightforward calculation, we obtain the profile likelihood function of ϕ_1, ϕ_2, ϕ_3 and λ:

$$\ell(\phi_1, \phi_2, \phi_3, \lambda) \propto \tilde{\sigma}_V(\phi_1, \phi_2, \phi_3, \lambda)^{-np} (\det \Sigma)^{-n/2}\ |J|.$$

By maximizing $\ell(\phi_1, \phi_2, \phi_3, \lambda)$, the maximum likelihood estimates $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\phi}_3$ and $\hat{\lambda}$ of ϕ_1, ϕ_2, ϕ_3 and λ can be derived; hence, the maximum likelihood estimates $\hat{\beta} = \tilde{\beta}(\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3, \hat{\lambda})$ and $\hat{\sigma}_V^2 = \tilde{\sigma}_V^2(\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3, \hat{\lambda})$ for β and σ_V^2 follow.

The conditional prediction of y_j, given Y, is now introduced as follows. By (1.3), through a straightforward calculation, the conditional density function of the Box-Cox power transformed variable $y_j^{(\lambda)}$, given β, σ_V^2, ϕ_1, ϕ_2, ϕ_3, λ and Y, can be expressed by

$$p(y_j^{(\lambda)}|\beta,\sigma_V^2,\phi_1,\phi_2,\phi_3,\lambda,Y) \propto (\det\ \sigma_V^2\Lambda_{22\cdot1})^{-1/2} \times \exp[-1/(2\sigma_V^2)\{y_j^{(\lambda)}-\mu_{2\cdot1}^{(\lambda)}\}^T\Lambda_{22\cdot1}^{-1}\{y_j^{(\lambda)}-\mu_{2\cdot1}^{(\lambda)}\}]. \quad (2.2)$$

Here, $\Lambda_{22\cdot1}=\Lambda_{22}-\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}$, $\mu_{2\cdot1}^{(\lambda)}=x\beta+\Lambda_{21}\Lambda_{11}^{-1}\{Y_j^{(\lambda)}-X\beta\}$, Λ_{11}, Λ_{12}, Λ_{21} and Λ_{22} are $p\times p$, $p\times 1$, $1\times p$ and 1×1 matrices, respectively, defined by

$$Cov\begin{pmatrix} Y_j^{(\lambda)} \\ y_j^{(\lambda)}\end{pmatrix} = \sigma_V^2(\phi_1\ \mathbf{1}\ \mathbf{1}^T + C + \phi_2\ I) = \sigma_V^2\ \Lambda = \sigma_V^2\ \begin{pmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{pmatrix},$$

$\mathbf{1}$ is a $(p+1)\times 1$ vector of 1's, I is a $(p+1)\times(p+1)$ identity matrix, and $C=\left(\phi_3^{|i-j|}\right)$, for $i,\ j=1,\cdots,p+1$.

By (2.2), we have the conditional expectation of $y_j^{(\lambda)}$, given β, σ_V^2, ϕ_1, ϕ_2, ϕ_3, λ and Y:

$$E(y_j^{(\lambda)}|\beta,\sigma_V^2,\phi_1,\phi_2,\phi_3,\lambda,Y)=\mu_{2\cdot1}^{(\lambda)},$$

for each $j=1,\cdots,n$. A natural predictor $\hat{y}_j$ for y_j, given Y, obtained by the MLM is

$$\begin{aligned}\hat{y}_j &= (1+\hat{\lambda})^{1/\hat{\lambda}}, && \text{for } \hat{\lambda}\neq 0,\\ &= \exp(\hat{\mu}_{2\cdot1}), && \text{for } \hat{\lambda}=0, \end{aligned} \quad (2.3)$$

for each $j=1,\cdots,n$. Here, $\hat{\mu}_{2\cdot1}$ is $\mu_{2\cdot1}^{(\lambda)}$ with its ϕ_1, ϕ_2, ϕ_3 and λ replaced respectively by their maximum likelihood estimates $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\phi}_3$ and $\hat{\lambda}$.

2.2 The ABA

Given the model specified by (1.2) and (1.3), the following two different types of priors for β, σ_V^2, ϕ_1, ϕ_2, ϕ_3 and λ are used to predict y_j:

$$\Pi(\beta,\sigma_V^2,\phi_1,\phi_2,\phi_3,\lambda) \propto \sigma_V^{-2}, \quad (2.4)$$

$$\Pi(\beta,\sigma_V^2,\phi_1,\phi_2,\phi_3,\lambda) \propto \sigma_V^{-2}\ \pi(\phi_1)\ \pi(\phi_2), \quad (2.5)$$

where the function π is the inverse gamma distribution with the hyperparameters $\zeta>0$ and $\theta>0$. Specifically,

$$\pi(x)=IG(\zeta,\theta)\propto x^{-(\zeta+1)}\exp(-\theta/x).$$

In both (2.4) and (2.5), it is assumed that β, λ, σ_V^2, ϕ_1, ϕ_2 and ϕ_3 have independent prior distributions. In (2.4), no information is available for each parameter, and is therefore referred to as noninformative prior (Edwards, Lindman and Savage 1963, and Zellner and Tiao 1964). Conversely, in (2.5), the prior is composed of inverse gamma distributions (Gelfand, Hills, Racince-Poon and Smith 1990). Note that the priors for ϕ_1 and ϕ_2 given in the right-hand side of (2.5) can be generalized by using different sets of hyperparameters in their inverse gamma distributions. Here, for the sake of simplicity, the same set of hyperparameters is given for the two inverse gamma distributions.

In order to employ the prior distribution specified in (2.5), Rissanen (1986) and Lee and Tsao (1993) suggest using the minimum accumulated prediction error (MAPE) criterion to choose the values of the hyperparameters ζ and θ of the inverse gamma distributions. Accordingly, the authors take the minimizer $\hat{\zeta}$ and $\hat{\theta}$ of the function $S(\zeta, \theta)$ over both ζ and θ as the selected values of ζ and θ, respectively. For the given values of ζ and θ, the accumulated prediction error $S(\zeta, \theta)$ of the corresponding ABA is defined by

$$S(\zeta, \theta) = \sum_{j=1}^{n} \sum_{k=4}^{p} |Y_{jk} - \hat{Y}_{jk}(\zeta, \theta)|.$$

Here, $\hat{Y}_{jk}(\zeta, \theta)$, the predicted value of Y_{jk}, is $\hat{y}_j$ in (2.7) when $p = k - 1$ is used in (1.3), and ζ and θ are employed in (2.5). Recall that y_j is the future observation to be observed at time $p + 1$. To compute its estimate $\hat{y}_j$, the values of the six parameters, including β_1 and β_2, need to be estimated by using the np observations made at equally spaced times $1, \cdots, p$ on the n subjects. To make the estimates for these linear regression parameters β_1 and β_2 with more meaning, the minimum value of p is taken as 3 by Rissanen (1986) and Lee and Tsao (1993); hence, the value of the subindex k in $S(\zeta, \theta)$ starts at 4. In this paper, the prior given in (2.4) is referred to as prior 1 and that in (2.5) with the selected hyperparameters $(\hat{\zeta}, \hat{\theta})$ is prior 2.

The conditional prediction of y_j, given Y, is now considered. By applying the approximate method in Ljung and Box (1980), through a straightforward calculation, we have the approximate predictive distribution of $y_j^{(\lambda)}$, given Y:

$$y_j^{(\lambda)}|Y \ \sim \ T_1[\hat{\mu}_y, \hat{S}_y\{(np-2)\hat{G}_{22}\}^{-1}, np-2]. \tag{2.6}$$

Here, $T_1(\mu, \sigma, q)$ stands for the distribution of the univariate random variable U such that $(U - \mu)/\sigma$ has a Student's t distribution with q degrees of freedom (Dickey 1967), and $\hat{\mu}_y$, $\hat{S}_y$ and $\hat{G}_{22}$ are μ_y, S_y and G_{22} with their ϕ_1, ϕ_2, ϕ_3 and

λ replaced respectively by $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\phi}_3$ and $\hat{\lambda}$ which maximize

$$p(\phi_1,\phi_2,\phi_3,\lambda|Y,\text{prior 1}) \propto (\det\Sigma)^{-(n-1)/2} \times (\det\Lambda \det Q \det G_{22})^{-1/2}\ S_y^{-(np-2)/2}\ |J|,$$

$$p(\phi_1,\phi_2,\phi_3,\lambda|Y,\text{prior 2}) \propto (\phi_1\phi_2)^{-(\tilde{\zeta}+2-1)}\ \exp(-\tilde{\theta}/\phi_1 - \tilde{\theta}/\phi_2) \times (\det\Sigma)^{-(n-1)/2}\ (\det\Lambda \det Q \det G_{22})^{-1/2}\ S_y^{-(np-2)/2}\ |J|.$$

Moreover,

$$\mu_y = x\beta_j^* - G_{22}^{-1}G_{21}\{Y_j^{(\lambda)} - X\beta_j^*\},$$

$$S_y = \sum_{\ell=1,\ell\neq j}^{n} \{Y_\ell^{(\lambda)} - X\beta_j^*\}^T\Sigma^{-1}\{Y_\ell^{(\lambda)} - X\beta_j^*\} + \{Y_j^{(\lambda)} - X\beta_j^*\}^T G_{11\cdot 2}\{Y_j^{(\lambda)} - X\beta_j^*\},$$

$$Q_1 = (n-1)X^T\Sigma^{-1}X, \quad Q_2 = \tilde{X}^T\Lambda^{-1}\tilde{X},$$

$$\tilde{X} = \binom{X}{x} = \begin{pmatrix} 1,1,\cdots,1 \\ 1,2,\cdots,p+1 \end{pmatrix}^T, \quad \beta_j^* = Q_1^{-1}\{\sum_{\ell=1,\ell\neq j}^{n} X^T\Sigma^{-1}Y_\ell^{(\lambda)}\},$$

$$G = \Lambda^{-1}\tilde{X}Q_2^{-1}Q_1(Q_1+Q_2)^{-1}\tilde{X}^T\Lambda^{-1} + Z(Z^T\Lambda Z)^{-1}Z^T = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

$$G_{11\cdot 2} = G_{11} - G_{12}G_{22}^{-1}G_{21},$$

where Z is a $(p+1)\times(p+1-2)$ matrix satisfying $\tilde{X}^TZ = 0$, and G_{11}, G_{12}, G_{21} and G_{22} are $p\times p$, $p\times 1$, $1\times p$ and 1×1 matrices, respectively.

By (2.6), a natural approximate predictor for $y_j^{(\lambda)}$ is $\hat{\mu}_y$, the expectation of the approximate predictive distribution of $y_j^{(\lambda)}$, given Y. Hence, the predictor $\hat{y}_j$ for y_j produced by the ABA with each prior 1 and 2 can be expressed by

$$\begin{aligned} \hat{y}_j &= (1+\hat{\lambda}\ \hat{\mu}_y)^{1/\hat{\lambda}}, && \text{for } \hat{\lambda}\neq 0, \\ &= \exp(\hat{\mu}_y), && \text{for } \hat{\lambda}=0 \end{aligned} \tag{2.7}$$

for each $j = 1, \cdots, n$.

2.3 Both the MLM and the ABA for the model secified by (1.2) and (1.4)

Given the model described by (1.2) and (1.4), through a straightforward calculation, the predictor $\hat{y}_j$ for y_j, given Y, obtained by the MLM is

$$\begin{aligned} \hat{y}_j &= (1+\lambda^*\ \mu_{2\cdot 1}^*)^{1/\lambda^*}, && \text{for } \lambda^*\neq 0, \\ &= \exp(\mu_{2\cdot 1}^*), && \text{for } \lambda^*=0, \end{aligned} \tag{2.8}$$

for each $j = 1, \cdots, n$. Here, $\mu^*_{2\cdot1}$ is $\hat{\mu}_{2\cdot1}$ in (2.3) derived both by replacing $Y_j^{(\lambda)}$ with $Y_j^{(\lambda^*)}$ and by deleting the random variable λ in Subsection 2.1. Also, given the value of λ^*, the predictor $\hat{y}_j$ for y_j, given Y, produced by the ABA with each prior 1 and 2 can be expressed by

$$\begin{aligned} \hat{y}_j &= (1 + \lambda^* \ \mu^*_y)^{1/\lambda^*}, && \text{for } \lambda^* \neq 0, \\ &= \exp(\mu^*_y), && \text{for } \lambda^* = 0, \end{aligned} \tag{2.9}$$

for each $j = 1, \cdots, n$. Here, μ^*_y is $\hat{\mu}_y$ in (2.7) derived both by replacing $Y_j^{(\lambda)}$ with $Y_j^{(\lambda^*)}$ and by deleting the random variable λ in Subsection 2.2.

3. Examples

Empirical studies were carried out as a means of obtaining further insight into the results of Section 2. In this section, models specified by (1.1) and (1.2), by (1.2) and (1.3), and by (1.2) and (1.4) are referred to as models 1, 2 and 3, respectively. Given each model 1, 2 and 3, the three prediction methods described in Section 2, MLM, ABA with prior 1 and ABA with prior 2, were applied to two real data examples. The performance of each prediction method was measured by both the mean absolute difference (MAD) and the mean absolute relative difference (MARD), whereas

$$MAD = n^{-1} \sum_{j=1}^{n} |\hat{y}_j - y_j|, \quad MARD = n^{-1} \sum_{j=1}^{n} |\hat{y}_j / y_j - 1|.$$

Example 3.1. The first data set consists of the weights of 30 calves, each being observed from 0 to 18 weeks in increments of 2 weeks. This data set is given in Group B of Table 6.1 of Diggle, Liang and Zeger (1994) with the weights on the 19th week excluded. The weights of the calves on the 18th week were predicted by using the data collected in the first 16 weeks. To avoid "overflow" or "underflow" problems for numerical computation, the weights of the calves were divided by 100. The resulting data are plotted in Figure 1a.

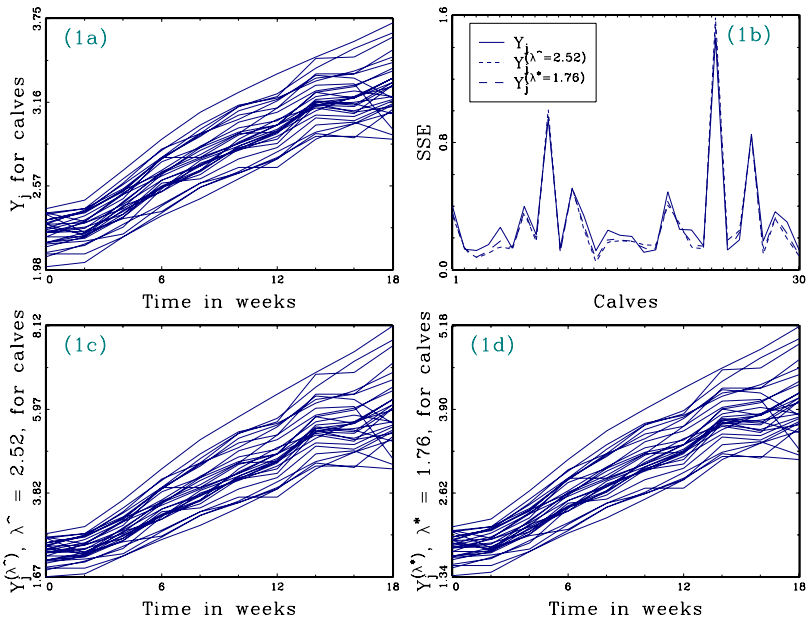

Figure 1: Plot of the original weights of 30 calves (1a), their corresponding power transformed values produced by the ABA with prior 2 for model 2 (1c) and by model 3 (1d), and the SSE (1b) derived by fitting the standardized linear regression model to the data for each calf in (1a), (1c), and (1d).

The data shown in Figure 1a were fitted to model 1 without applying the Box-Cox power transformation. For this, the data shown in Figure 1a were examined, and were shown to have a linear trend. By this and the equally spaced times of measurement, the vector of regression coefficients was taken to be $\beta = (\beta_1, \beta_2)^T$, and the design matrix was taken to be $X = \begin{pmatrix} 1,1,\cdots,1 \\ 1,2,\cdots,9 \end{pmatrix}^T$ for each calf. To employ prior 2, the MAPE criterion was used to select the inverse gamma hyperparameters ζ and θ over the rectangle $(1, 40] \times (0, 10]$. The values of $S(\zeta, \theta)$ were calculated on 1950×500 equally spaced grid points in the rectangle. The minimizer of these values occurred at $(\hat{\zeta}, \hat{\theta}) = (28.7, 1.5)$, and was taken to be the selected value of (ζ, θ) for prior 2. Then, the weights of the calves on the 18th week were predicted using each of the above three methods.

Similar computation procedures were also performed to fit the data shown in Figure 1a to each model 2 and 3 with the Box-Cox power transformation. Here, the selected values of $(\hat{\zeta}, \hat{\theta})$ for prior 2 were $(1.05, 0.30)$ and $(2.55, 1.78)$ for these two models, respectively. The estimated values $\hat{\lambda}$ of the Box-Cox power transformation parameter λ were 0.96, 1.12 and 2.52 obtained respectively by the three prediction methods for model 2. The value of λ^* was 1.76 for model 3. Note that the values of $\hat{\lambda}$ derived by both the MLM and the ABA with prior 1

Table 1: Comparison of prediction accuracy of y_j for the weights of 30 calves.

		MLM	Prior 1	Prior 2
Model 1				
	MAD	0.0894	0.0889	0.0829
	MARD	0.0287	0.0285	0.0266
Model 2				
	MAD	0.0898	0.0879	0.0771
	MARD	0.0288	0.0282	0.0246
Model 3				
	MAD	0.0813	0.0810	0.0752
	MARD	0.0260	0.0259	0.0232

for model 2 are close to 1; hence, the resulting Box-Cox power transformed data are nearly the same as the original data. Judging from this, it is expected that the prediction abilities of these two methods would not be significantly improved by using the Box-Cox power transformation. On the other hand, Figures 1c and 1d show the Box-Cox power transformed data of Figure 1a as produced by the ABA with prior 2 for model 2, as well as obtained by model 3, respectively.

Because the main focus of the study was to ascertain whether the power transformed data in Figures 1c and 1d have better linear fit than the original data in Figure 1a, the data shown by each curve in these figures were fitted to the standardized linear regression model, and their corresponding error sum of squares, denoted by SSE, was calculated. Here, the SSE represents the measurement of the distance between the given curve and its associated straight line obtained by the linear regression model. The lower the value of the SSE, the better the linear fit. Figure 1b shows the SSE for each curve in Figures 1a, 1c and 1d. The sample 25th, 50th and 75th percentiles of these SSE for the 30 calves were (0.130,0.233,0.370), (0.100,0.198,0.323) and (0.104,0.202,0.330) for the data in Figures 1a, 1c and 1d, respectively. According to the results, these power transformed data exhibit better linear fit than the original data. It is also expected that the prediction abilities of the methods corresponding to Figures 1c and 1d would be better.

The numerical results for the weights of the calves are given in Table 1. Table 1 contains the MAD and MARD of the predicted values obtained by the three prediction methods using each model 1, 2 and 3. It shows that the prediction abilities of the three methods for model 1 can be significantly improved by using power transformation, in the sense that they have lower levels of both MAD and MARD. For example, using the ABA with prior 2, 7.00% and 9.29% of the value of MAD for model 1 without power transformation can be reduced by simply

Table 2: Weights (with unit as 10 kg) of 18 students.

	Time in semesters										
#	1	2	3	4	5	6	7	8	9	10	11
1	2.04	2.20	2.18	2.40	2.50	2.70	2.92	3.00	3.50	3.80	4.08
2	1.92	2.22	2.22	2.45	2.58	2.70	2.92	3.04	3.02	3.44	3.44
3	1.96	2.12	2.16	2.34	2.40	2.50	2.62	2.98	3.36	3.62	3.72
4	2.00	2.40	2.50	2.90	3.20	3.60	3.65	4.06	4.12	4.26	4.30
5	1.90	1.94	2.08	2.32	2.42	2.62	3.02	3.10	3.46	3.72	3.92
6	2.02	2.20	2.16	2.46	2.38	2.56	2.62	2.94	3.50	3.85	4.02
7	2.46	2.58	2.60	2.90	3.30	3.20	3.76	3.70	4.34	4.46	4.90
8	2.50	2.62	2.80	3.01	3.32	3.48	3.90	4.10	4.44	4.90	4.98
9	1.52	1.60	1.54	1.72	1.84	1.88	2.08	2.16	2.22	2.40	2.50
10	2.54	2.60	2.74	3.02	3.64	4.04	4.40	4.52	4.88	5.20	5.90
11	3.00	3.26	3.60	3.60	4.10	4.10	4.40	4.60	4.80	4.88	5.30
12	1.94	2.02	2.10	2.40	2.36	2.58	2.78	2.80	3.04	3.14	3.36
13	2.68	2.90	3.32	3.55	3.84	4.02	4.44	4.50	5.20	5.40	5.60
14	2.32	2.44	2.50	2.80	2.94	3.06	3.48	3.58	4.50	4.58	4.82
15	1.82	2.00	2.10	2.30	2.30	2.40	2.62	2.70	2.94	3.02	3.08
16	2.50	2.56	2.96	3.10	3.46	3.88	4.18	4.66	5.14	5.70	6.04
17	2.18	2.38	2.44	2.70	2.70	2.86	3.00	3.20	3.24	3.50	3.56
18	2.18	2.40	2.38	2.60	2.74	3.00	3.20	3.28	3.40	3.60	4.30

employing models 2 and 3 with power transformation, respectively. A similar conclusion can be drawn for the value of MARD. Table 1 also demonstrates that the ABA with prior 2 has the best prediction ability, regardless of whether or not power transformation is performed. This is evidenced by a significant improvement in both MAD and MARD. This advantage that the ABA with prior 2 has better prediction ability than the ABA with prior 1 might be due to the fact that prior 2 provides more information about the parameters than prior 1.

Example 3.2. The second data set consists of the weights of 18 students in the same class at a primary school operating the government-provided lunch program in Hualien, Taiwan, R.O.C. The weight of each student was measured at the beginning of each semester. According to government regulations, this data set was collected with the purpose of monitoring the weight growth status of each student so that physical drawbacks could be discovered early, and so that proper teaching activities could be provided to eliminate such drawbacks. Such regulations can be found at the URL address http://www.edu.tw/physical/rules/0724-2.htm. In this project, the weights of the students in the 11th semester were predicted by using those obtained in the first 10 semesters. To avoid "overflow" or "underflow"

problems in numerical computation, the weights of the students were divided by 10. The resulting data for these students are plotted in Figure 2a and listed in Table 2.

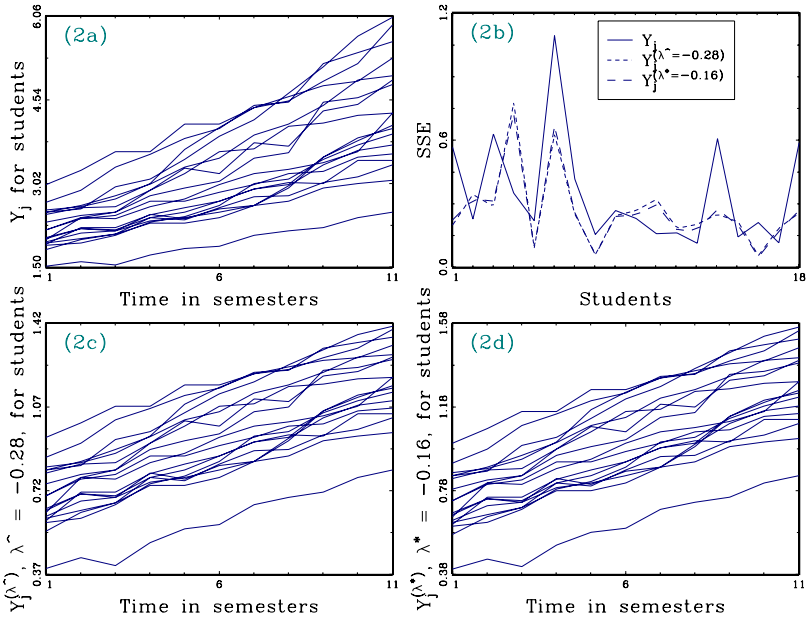

Figure 2: Plot of the original weights of 18 students (2a), their corresponding power transformed values produced by the ABA with prior 2 for model 2 (2c) and by model 3 (2d), and the SSE (2b) derived by fitting the standardized linear regression model to the data for each student in (2a), (2c), and (2d).

Given each model 1, 2 and 3, the three prediction methods were applied to the data in Example 3.2. The numerical results are now presented. By both the linear trend of the data presented in Figure 2a and the equally spaced times of measurement, the vector of regression coefficients was taken to be $\beta = (\beta_1, \beta_2)^T$ and the design matrix was taken to be $X = \begin{pmatrix} 1,1,\cdots,1 \\ 1,2,\cdots,10 \end{pmatrix}^T$ for each student. Given each model, to select the values of ζ and θ for prior 2, the values of $S(\zeta, \theta)$ were evaluated on 1950×500 equally spaced grid points in the rectangle $(1, 40] \times (0, 10]$. For each model 1, 2 and 3, the selected value of $(\hat{\zeta}, \hat{\theta})$ for prior 2 was $(30, 1.58)$, $(4.55, 0.03)$ and $(5, 0.11)$, respectively. On the other hand, the estimated values $\hat{\lambda}$ of the power transformation parameter λ were 0.57, 0.49 and -0.28 obtained respectively by the three prediction methods for model 2. The value of λ^* was -0.16 for model 3.

Figures 2c and 2d show the power transformed data of Figure 2a produced by the ABA with prior 2 for model 2, as well as obtained by model 3, respectively. Comparing visually the curves in Figures 2a, 2c and 2d, it is clear that

Table 3: Comparison of prediction accuracy of y_j for the weights of 18 students.

		MLM	Prior 1	Prior 2
Model 1				
	MAD	0.1762	0.1755	0.1758
	MARD	0.0489	0.0487	0.0489
Mode 2				
	MAD	0.1699	0.1664	0.1654
	MARD	0.0404	0.0396	0.0393
Mode 3				
	MAD	0.1557	0.1555	0.1559
	MARD	0.0375	0.0374	0.0378

those in Figure 2a have larger curvature. Figure 2b contains the SSE obtained by fitting the data presented by each curve in Figures 2a, 2c and 2d to the standardized linear regression model. The sample 25th, 50th and 75th percentiles of these SSE for the 18 students were (0.151,0.266,0.499), (0.137,0.231,0.335) and (0.134,0.228,0.333) for the data in Figures 2a, 2c and 2d, respectively. By the results, we conclude that the power transformed data in Figures 2c and 2d have better linear fit than the original data in Figure 2a.

Table 3 contains the MAD and MARD of the predicted values obtained by the three prediction methods using each model 1, 2 and 3. Considering the values of both MAD and MARD, Table 3 shows that the prediction abilities of the three methods for model 1 can be significantly improved by using the Box-Cox power transformation. For example, using the ABA with prior 2, 5.92% and 11.32% of the value of MAD for model 1 without power transformation can be reduced by employing models 2 and 3 with power transformation, respectively.

References

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformation. *J. Roy. Statist. Soc. Ser. A* **26**, 211-252.

Dickey, J. M. (1967). Matricvariate generalizations of the multivariate distribution and inverted multivariate t distribution. *Ann. Math. Stat.* **38**, 511-518.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959-971.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford Statistical Science Series #13, Oxford Science Publications.

Donnelly, C. A., Laird, N. M. and Ware, J. H. (1995). Prediction and creation of smooth curve for temporally correlated longitudinal data. *J. Amer. Statist. Assoc.* **90**,

984-989.

Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193-242.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. S. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.* **85**, 972-985.

Lee, J. C. (1988). Prediction and estimation of growth curve with special covariance structures. *J. Amer. Statist. Assoc.* **83**, 432-440.

Lee, J. C. and Tsao, S. L. (1993). On estimation and prediction procedures for AR(1) models with power transformation. *J. Forecast.* **12**, 499-511.

Ljung, G. M. and Box, G. E. P. (1980). Analysis of variance with autocorrelated observations. *Scand. J. Statist.* **7**, 172-180.

Neter, J., Wasserman, W. and Kutner, M. H. (1989). *Applied Linear Regression Models.* Richard D. Irwin.

Rissanen, J. (1986). Order estimation by accumulated prediction errors. In *Essays in Time Series and Allied Processes — Papers in Honor of E.J. Hannan* (Edited by Gani, J. and Priestley, M. B.), Sheffield.

Seber, G. A. F. (1977). *Linear Regression Analysis.* John Wiley.

Zellner, A. and Tiao, G. C. (1964). Bayesian analysis of regression model with autocorrelated errors. *J. Amer. Statist. Assoc.* **59**, 763-778.

Received Octomber 3, 2002; accepted April 1, 2003.

R. C. Hwang
Department of Accounting and Statistics
Dahan Institute of Technology
Hualien, Taiwan 971, ROC
sunny@ms01.dahan.edu.tw