International Journal of Pure and Applied Mathematics

Volume 5 No. 2 2003, 159-172

ON REGRESSION MODELS WITH AUTOCORRELATED ERROR: SMALL SAMPLE PROPERTIES

Hisashi Tanizaki

Graduate School of Economics
Kobe University
2-1, Rokkodai-cho, Nada-ku
Kobe 657-8501, JAPAN
e-mail: tanizaki@kobe-u.ac.jp

Abstract: Using both the maximum likelihood estimator and the Bayes estimator, we consider estimating the regression model with the first-order autocorrelated error, where the initial distribution of the autocorrelated error is taken into account. For the Bayes estimator, the Gibbs sampler and the Metropolis-Hastings algorithm are utilized to obtain random draws of the parameters. As a result, the Bayes estimator is less biased and more efficient than the maximum likelihood estimator. Especially, for the autocorrelation coefficient, the Bayes estimate is much less biased than the maximum likelihood estimate. Accordingly, for the standard error of the estimated regression coefficient, the Bayes estimate is more plausible than the maximum likelihood estimate, because variance of the estimated regression coefficient depends on the estimated autocorrelation coefficient. Thus, we find that the Bayes approach might be recommended in the empirical studies.

AMS Subject Classification: 62M10, 62F15
Key Words: autoregressive model, Gibbs sampler, Metropolis-Hastings algorithm, Bayes estimator, MLE

1. Introduction

In this paper, we consider the regression model with the first-order autocorrelated error term, where the error term is assumed to be stationary, i.e., the

Received: February 4, 2003

autocorrelation coefficient is assumed to be less than one in absolute value. The traditional estimator, i.e., the maximum likelihood estimator (MLE), is compared with the Bayes estimator (BE). Utilizing the Gibbs sampler, Chib [1] and Chib and Greenberg [2] discussed the regression model with the autocorrelated error term in a Bayesian framework, where the initial condition of the autoregressive process is ignored. In this paper, taking into account the initial density, we compare MLE and BE, where the Gibbs sampler and the Metropolis-Hastings (MH) algorithm are utilized in BE. As for MLE, it is well known that the autocorrelation coefficient is underestimated in small sample and therefore that variance of the estimated regression coefficient is biased (see, for example, Andrews [3] and Tanizaki [4, 5]). Under this situation, inference on the regression coefficient is not appropriate. We show in this paper that BE is superior to MLE, because BE of both the autocorrelation coefficient and the variance of the error term are closer to the true values, compared with MLE's.

2. Setup of the Model

Let X_t be a $1 \times k$ vector of exogenous variables and β be a $k \times 1$ parameter vector. Consider the following standard linear regression model:

$$y_t = X_t\beta + u_t, \qquad u_t = \rho u_{t-1} + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2),$$

for $t = 1, 2, \cdots, n$, where ϵ_1, ϵ_2, $\cdots$, ϵ_n are assumed to be mutually independently distributed. In this model, the parameter to be estimated is given by $\theta \equiv (\beta', \rho, \sigma^2)'$.

The unconditional density function of y_t is:

$$f(y_t|\beta, \rho, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2/(1-\rho^2)}} \exp\Big(-\frac{1}{2\sigma^2/(1-\rho^2)}(y_t - X_t\beta)^2\Big),$$

which corresponds to the initial density function of y_t when $t = 1$. Let Y_t be the information set up to time t, i.e., $Y_t = \{y_t, y_{t-1}, \cdots, y_1\}$. The conditional density of y_t given Y_{t-1} is:

$$\begin{aligned} f(y_t|Y_{t-1}, \beta, \rho, \sigma^2) &= f(y_t|y_{t-1}, \beta, \rho, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}\big((y_t - \rho y_{t-1}) - (X_t - \rho X_{t-1})\beta\big)^2\Big). \end{aligned}$$

Therefore, the joint density of Y_n, i.e., the likelihood function, is given by:

$$f(Y_n|\beta,\rho,\sigma^2) = f(y_1|\beta,\rho,\sigma^2)\prod_{t=2}^{n} f(y_t|Y_{t-1},\beta,\rho,\sigma^2)$$
$$= (2\pi\sigma^2)^{-n/2}(1-\rho^2)^{1/2}\exp\Big(-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t^* - X_t^*\beta)^2\Big), \quad (1)$$

where y_t^* and X_t^* represent the following transformed variables:

$$y_t^* \equiv y_t^*(\rho) = \begin{cases} \sqrt{1-\rho^2}y_t, & \text{for } t=1, \\ y_t - \rho y_{t-1}, & \text{for } t = 2,3,\cdots,n, \end{cases}$$

$$X_t^* \equiv X_t^*(\rho) = \begin{cases} \sqrt{1-\rho^2}X_t, & \text{for } t=1, \\ X_t - \rho X_{t-1}, & \text{for } t = 2,3,\cdots,n, \end{cases}$$

which depend on the autocorrelation coefficient ρ.

2.1. Maximum Likelihood Estimator

As shown above, the likelihood function is given by (1). Maximizing (1) with respect to β and σ^2, we obtain the following expressions:

$$\hat{\beta} \equiv \hat{\beta}(\rho) = (\sum_{t=1}^{n} X_t^{*\prime}X_t^*)^{-1}\sum_{t=1}^{n} X_t^{*\prime}y_t^*, \quad (2)$$

$$\hat{\sigma}^2 \equiv \hat{\sigma}^2(\rho) = \frac{1}{n}\sum_{t=1}^{n}(y_t^* - X_t^*\hat{\beta})^2. \quad (3)$$

By substituting $\hat{\beta}$ and $\hat{\sigma}^2$ into β and σ^2 in (1), we have the concentrated likelihood function:

$$f(Y_n|\hat{\beta},\rho,\hat{\sigma}^2) = \Big(2\pi\hat{\sigma}^2(\rho)\Big)^{-n/2}(1-\rho^2)^{1/2}\exp\big(-\frac{n}{2}\big), \quad (4)$$

which is a function of ρ. (4) is maximized with respect to ρ. In the next section, we obtain the maximum likelihood estimate of ρ by a simple grid search, in which the concentrated likelihood function (4) is maximized by changing the parameter value of ρ by 0.0001 in the interval between -0.9999 and 0.9999. Once the solution of ρ, denoted by $\hat{\rho}$, is obtained, $\hat{\beta}(\hat{\rho})$ and $\hat{\sigma}^2(\hat{\rho})$ lead to the maximum likelihood estimates of β and σ^2. Hereafter, $\hat{\beta}$, $\hat{\sigma}^2$ and $\hat{\rho}$ are taken

as the maximum likelihood estimates of β, σ^2 and ρ, i.e., $\hat{\beta}(\hat{\rho})$ and $\hat{\sigma}^2(\hat{\rho})$ are simply written as $\hat{\beta}$ and $\hat{\sigma}^2$.

Variance of the estimate of $\theta = (\beta', \sigma^2, \rho)'$ is asymptotically given by: $\mathrm{V}(\hat{\theta}) = I^{-1}(\theta)$, where $I(\theta)$ denotes the information matrix, which is represented as:

$$I(\theta) = -\mathrm{E}\left(\frac{\partial^2 \log f(Y_n|\theta)}{\partial\theta\partial\theta'}\right).$$

Therefore, the variance of $\hat{\beta}$ is given by $\mathrm{V}(\hat{\beta}) = \sigma^2(\sum_{t=1}^n X_t^{*\prime} X_t^*)^{-1}$ in large sample, where ρ in X_t^* is replaced by $\hat{\rho}$, i.e., $X_t^* = X_t^*(\hat{\rho})$. For example, suppose that X_t^* has a tendency to rise over time t and that we have $\rho > 0$. If ρ is underestimated, then $\mathrm{V}(\hat{\beta})$ is also underestimated, which yields incorrect inference on the regression coefficient β. Thus, unless ρ is properly estimated, the estimate of $\mathrm{V}(\hat{\beta})$ is also biased. In large sample, $\hat{\rho}$ is a consistent estimator of ρ and therefore $\mathrm{V}(\hat{\beta})$ is not biased. However, in small sample, since it is known that $\hat{\rho}$ is underestimated (see, for example, Andrews [3], Tanizaki [4, 5]), clearly $\mathrm{V}(\hat{\beta})$ is also underestimated. In addition to $\hat{\rho}$, the estimate of σ^2 also influences inference of β, because we have $\mathrm{V}(\hat{\beta}) = \sigma^2(\sum_{t=1}^n X_t^{*\prime} X_t^*)^{-1}$ as mentioned above. If σ^2 is underestimated, the estimated variance of β is also underestimated. $\hat{\sigma}^2$ is a consistent estimator of σ^2 in large sample, but it is appropriate to consider that $\hat{\sigma}^2$ is biased in small sample, because $\hat{\sigma}^2$ is a function of $\hat{\rho}$ as in (3). Therefore, the biased estimate of ρ gives us the serious problem on inference of β.

2.2. Bayes Estimator

We assume that the prior density functions of β, ρ and σ^2 are the following noninformative priors:

$$f_\beta(\beta) \propto \text{const.}, \qquad f_\rho(\rho) \propto \text{const.}, \qquad f_\sigma(\sigma^2) \propto \frac{1}{\sigma^2}, \tag{5}$$

for $-\infty < \beta < \infty$, $-1 < \rho < 1$ and $0 < \sigma < \infty$. For $f_\rho(\rho)$, theoretically we should have $-1 < \rho < 1$. As for the prior density of σ^2, since we consider that $\log\sigma^2$ has the flat prior for $-\infty < \log\sigma^2 < \infty$, we obtain $f_\sigma(\sigma^2) \propto 1/\sigma^2$.

Combining the four densities (1) and (5), the posterior density function of β, ρ and σ^2, denoted by $f_{\beta\rho\sigma}(\beta, \rho, \sigma^2|Y_n)$, is represented as follows:

$$\begin{aligned} f_{\beta\rho\sigma}(\beta, \rho, \sigma^2|Y_n) &\propto f(Y_n|\beta, \rho, \sigma^2) f_\beta(\beta) f_\rho(\rho) f_\sigma(\sigma^2) \\ &\propto (\sigma^2)^{-(n/2+1)}(1-\rho^2)^{1/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{t=1}^n (y_t^* - X_t^*\beta)^2\Big). \end{aligned} \tag{6}$$

We consider generating random draws of β, ρ and σ^2 given Y_n. However, it is not easy to generate random draws of β, ρ and σ^2 from $f_{\beta\rho\sigma}(\beta, \rho, \sigma^2|Y_n)$. Therefore, we perform the Gibbs sampler in this problem. According to the Gibbs sampler, we can sample from the posterior density function (6), using the three conditional distributions $f_{\beta|\rho\sigma}(\beta|\rho, \sigma^2, Y_n)$, $f_{\rho|\beta\sigma}(\rho|\beta, \sigma^2, Y_n)$ and $f_{\sigma|\beta\rho}(\sigma^2|\beta, \rho, Y_n)$, which are proportional to $f_{\beta\rho\sigma}(\beta, \rho, \sigma^2|Y_n)$ and are obtained as follows:

$$f_{\beta|\rho\sigma}(\beta|\rho, \sigma^2, Y_n) \propto \exp\Big(-\frac{1}{2}(\beta - \hat{\beta})'(\frac{1}{\sigma^2}\sum_{t=1}^{n} X_t^{*\prime} X_t^*)(\beta - \hat{\beta})\Big), \tag{7}$$

$$f_{\rho|\beta\sigma}(\rho|\beta, \sigma^2, Y_n) \propto (1 - \rho^2)^{1/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{t=1}^{n}\Big(y_t^* - X_t^*\beta\Big)^2\Big), \tag{8}$$

$$f_{\sigma|\beta\rho}(\sigma^2|\beta, \rho, Y_n) \propto \frac{1}{(\sigma^2)^{n/2+1}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t^* - X_t^*\beta)^2\Big). \tag{9}$$

For (7), $\hat{\beta}$ represents the ordinary least squares (OLS) estimate, which is represented by $\hat{\beta} = (\sum_{t=1}^{n} X_t^{*\prime} X_t^*)^{-1} (\sum_{t=1}^{n} X_t^{*\prime} y_t^*)$. (7) indicates that β is sampled from the multivariate normal distribution:

$$\beta \sim N\big(\hat{\beta}, \sigma^2(\sum_{t=1}^{n} X_t^{*\prime} X_t^*)^{-1}\big).$$

(8) is not represented in a known distribution, where $-1 < \rho < 1$. Sampling from (8) is implemented by the MH algorithm, which will be discussed below. (9) indicates that $1/\sigma^2 \sim G(n/2,\ 2/\sum_{t=1}^{n} \epsilon_t^2)$, where $\epsilon_t = y_t^* - X_t^*\beta$. Note that $G(a, b)$ denotes the gamma distribution with parameters a and b (see, for example, Bernardo and Smith [6], Carlin and Louis [7], Chen et al [8], Gamerman [9], Robert and Casella [10] and Smith and Roberts [11] for the Gibbs sampler and the MH algorithm).

In order to generate random draws of β, ρ and σ^2 from the posterior density $f_{\beta\rho\sigma}(\beta, \rho, \sigma^2|Y_n)$, the following procedures have to be taken:

(i) Let β_i, ρ_i and σ_i^2 be the i-th random draws of β, ρ and σ^2. Take the initial values of $(\beta,\ \rho,\ \sigma^2)$ as $(\beta_{-M},\ \rho_{-M},\ \sigma_{-M}^2)$.

(ii) From $\beta \sim N\big(\hat{\beta},\ \sigma_{i-1}^2(\sum_{t=1}^{n} X_t^{*\prime} X_t^*)^{-1}\big)$ as in (7), generate β_i given ρ_{i-1}, σ_{i-1}^2 and Y_n, where $\hat{\beta} = (\sum_{t=1}^{n} X_t^{*\prime} X_t^*)^{-1}(\sum_{t=1}^{n} X_t^{*\prime} y_t^*)$, $y_t^* = y_t^*(\rho_{i-1})$ and $X_t^* = X_t^*(\rho_{i-1})$.

(iii) From (8), generate ρ_i given β_i, σ_{i-1}^2 and Y_n. Since it is not easy to generate random draws directly from (8), the MH algorithm is utilized, which is implemented as follows:

(a) Generate ρ^* from the uniform distribution between -1 and 1, which implies that the sampling density of ρ is given by $f_*(\rho|\rho_{i-1}) = 1/2$ for $-1 < \rho < 1$. Compute the acceptance probability $\omega(\rho_{i-1}, \rho^*)$, which is defined as:

$$\begin{aligned}\omega(\rho_{i-1}, \rho^*) &= \min\left(\frac{f_{\rho|\beta\sigma}(\rho^*|\beta_i, \sigma_{i-1}^2, Y_n)/f_*(\rho^*|\rho_{i-1})}{f_{\rho|\beta\sigma}(\rho_{i-1}|\beta_i, \sigma_{i-1}^2, Y_n)/f_*(\rho_{i-1}|\rho^*)},\ 1\right)\\ &= \min\left(\frac{f_{\rho|\beta\sigma}(\rho^*|\beta_i, \sigma_{i-1}^2, Y_n)}{f_{\rho|\beta\sigma}(\rho_{i-1}|\beta_i, \sigma_{i-1}^2, Y_n)},\ 1\right).\end{aligned}$$

(b) Set $\rho_i = \rho^*$ with probability $\omega(\rho_{i-1}, \rho^*)$ and $\rho_i = \rho_{i-1}$ otherwise.

(iv) From $1/\sigma^2 \sim G(n/2,\ 2/\sum_{t=1}^n \epsilon_t^2)$ as in (9), generate σ_i^2 given β_i, ρ_i and Y_n, where $\epsilon_t = y_t^* - X_t^*\beta$, $y_t^* = y_t^*(\rho_i)$ and $X_t^* = X_t^*(\rho_i)$.

(v) Repeat Steps (ii) – (iv) for $i = -M+1, -M+2, \cdots, N$, where M indicates the burn-in period.

Repetition of Steps (ii) – (iv) corresponds to the Gibbs sampler. For sufficiently large M, we have the following results:

$$\frac{1}{N}\sum_{i=1}^{N} g(x_i) \ \longrightarrow\ \mathrm{E}\big(g(x)\big),$$

where x should be replaced by β, ρ or σ^2. $g(\cdot)$ represents a function, typically $g(x) = x$ or $g(x) = x^2$. Thus, we can take the Bayes estimates of β, ρ and σ^2 as $\widetilde{\beta} \equiv (1/N)\sum_{i=1}^N \beta_i$, $\widetilde{\rho} \equiv (1/N)\sum_{i=1}^N \rho_i$ and $\widetilde{\sigma}^2 \equiv (1/N)\sum_{i=1}^N \sigma_i^2$, respectively.

3. Monte Carlo Experiments

For the exogenous variables, we take the data shown in Table 3, which are presented in Judge et al [12, p.156]. The DGP is defined as:

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + u_t, \qquad u_t = \rho u_{t-1} + \epsilon_t, \tag{10}$$

where the ϵ_t's are normally and independently distributed with $\mathrm{E}(\epsilon_t) = 0$ and $\mathrm{E}(\epsilon_t^2) = \sigma^2$. As in Judge et al [12], the parameter values are set to be $\beta' = (\beta_1,$

t	1	2	3	4	5	6	7	8	9	10
$x_{2,t}$	14.53	15.30	15.92	17.41	18.37	18.83	18.84	19.71	20.01	20.26
$x_{3,t}$	16.74	16.81	19.50	22.12	22.34	17.47	20.24	20.37	12.71	22.98
t	11	12	13	14	15	16	17	18	19	20
$x_{2,t}$	20.77	21.17	21.34	22.91	22.96	23.69	24.82	25.54	25.63	28.73
$x_{3,t}$	19.33	17.04	16.74	19.81	31.92	26.31	25.93	21.96	24.05	25.66

Table 1: The exogenous variables $x_{1,t}$ and $x_{2,t}$

β_2, $\beta_3) = (10, 1, 1)$. In this section, we utilize $x_{2,t}$ and $x_{3,t}$ given in Judge et al [12, p.156], which is shown in Table 1, and generate L samples of y_t given the $X_t = (1, x_{2,t}, x_{3,t})$ for $t = 1, 2, \cdots, n$. That is, we perform L simulation runs for both MLE and BE, where $L = 10^4$ is taken in this section.

The simulation procedure in this section is as follows:

(i) Given ρ, generate random numbers of u_t for $t = 1, 2, \cdots, n$, based on the assumptions: $u_t = \rho u_{t-1} + \epsilon_t$, $\epsilon_t \sim N(0, \sigma^2)$, where $\sigma^2 = 1$ and $\rho = -0.99$, -0.98, $\cdots$, 0.99 are taken.

(ii) Given β, X_t and u_t for $t = 1, 2, \cdots, n$, we obtain a set of data y_t, $t = 1, 2, \cdots, n$, from (10), where $(\beta_1, \beta_2, \beta_3) = (10, 1, 1)$ is assumed.

(iii) Given (y_t, X_t) for $t = 1, 2, \cdots, n$, obtain the estimates of $\theta = (\beta', \rho, \sigma^2)'$ by MLE and BE, which are denoted by $\hat{\theta}$ and $\widetilde{\theta}$, respectively.

(iv) Repeat (i) – (iii) L times, where $L = 10^4$ is taken as mentioned above.

(v) From L estimates of θ, compute the arithmetic average (AVE), the standard error (SER), the root mean square error (RMSE), the skewness (Skewness), the kurtosis (Kurtosis), and the 5, 10, 25, 50, 75, 90 and 95 percent points (5%, 10%, 25%, 50%, 75%, 90% and 95%) for each estimator. For AVE and RMSE of MLE, we compute:

$$\text{AVE} = \frac{1}{L}\sum_{l=1}^{L}\hat{\theta}_j^{(l)}, \qquad \text{RMSE} = \Big(\frac{1}{L}\sum_{l=1}^{L}(\hat{\theta}_j^{(l)} - \theta_j)^2\Big)^{1/2},$$

for $j = 1, 2, \cdots, 5$, where θ_j denotes the j-th element of θ and $\hat{\theta}_j^{(l)}$ represents the j-th element of $\hat{\theta}$ in the l-th simulation run. For AVE and RMSE of BE, simply $\hat{\theta}$ is replaced by $\widetilde{\theta}$.

In this section, we compare BE with MLE through Monte Carlo studies.

In Figure 3 we draw the relationship between ρ and $\hat{\rho}$, where $\hat{\rho}$ denotes the arithmetic average of the 10^4 MLE's, while in Figure 1 we display the

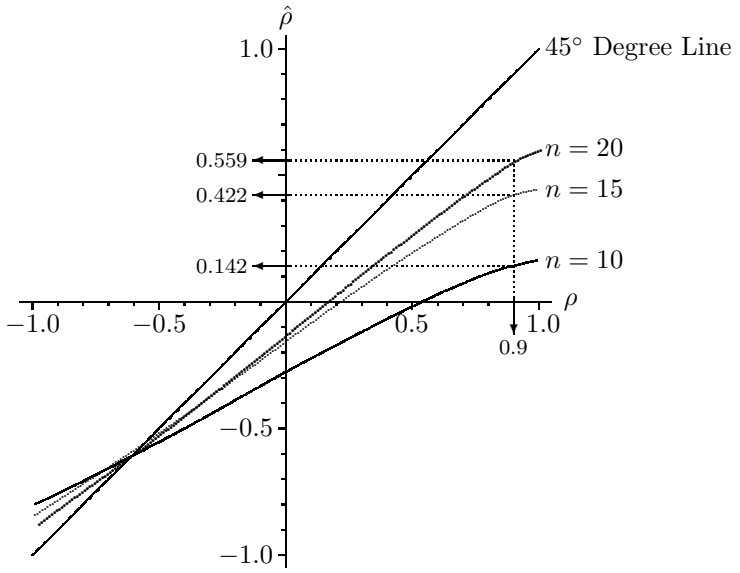

Figure 1: The arithmetic average from the 10^4 MLE's of AR(1) Coeff.

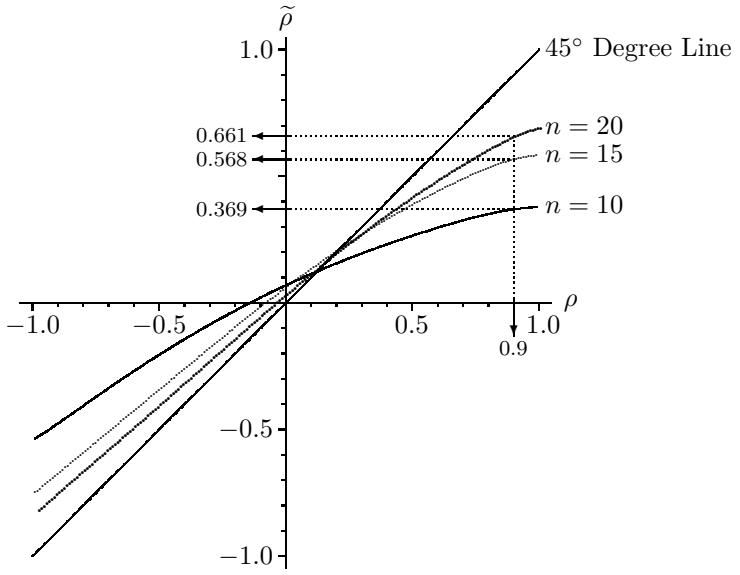

Figure 2: The arithmetic average from the 10^4 BE's of AR(1) Coeff.
——— $M = 5000$ and $N = 10^4$ ———

relationship between ρ and $\widetilde{\rho}$, where $\widetilde{\rho}$ indicates the arithmetic average of the 10^4 BE's. In the two figures the cases of $n = 10, 15, 20$ are shown, and $(M, N) = (5000, 10^4)$ is taken in Figure 1. In Appendix, we check whether M and N are large enough. If the relationship between ρ and $\hat{\rho}$ (or $\widetilde{\rho}$) lies on the 45° degree line, we can conclude that MLE (or BE) of ρ is unbiased. However, from the two figures, both estimators are biased. Take an example of $\rho = 0.9$ in Figure 3 and Figure 1. When the true value is $\rho = 0.9$, the arithmetic averages of 10^4 MLE's are given by 0.142 for $n = 10$, 0.422 for $n = 15$ and 0.559 for $n = 20$ (see Figure 3), while those of 10^4 BE's are 0.369 for $n = 10$, 0.568 for $n = 15$ and 0.661 for $n = 20$ (see Figure 1). As n increases the estimators are less biased, because MLE gives us the consistent estimators. Comparing BE and MLE, BE is less biased than MLE in the small sample, because BE is closer to the 45° degree line than MLE. Especially, as ρ goes to one, the difference between BE and MLE becomes quite large in small sample.

Table 3 and Table 2 represent the basic statistics such as arithmetic average, standard error, root mean square error, skewness, kurtosis and percent points, which are computed from $L = 10^4$ simulation runs, where the case of $n = 20$ and $\rho = 0.9$ is examined. Table 3 is based on the MLE's while Table 2 is obtained from the BE's.

Both MLE and BE give us the unbiased estimators of regression coefficients β_1, β_2 and β_3, because the arithmetic averages from the 10^4 estimates of β_1, β_2 and β_3, (i.e., AVE in the tables) are very close to the true parameter values, which are set to be $(\beta_1, \beta_2, \beta_3) = (10, 1, 1)$. However, in the SER and RMSE criteria, BE is better than MLE, because SER and RMSE of BE are smaller than those of MLE. From Skewness and Kurtosis in the two tables, we can see that the empirical distributions of MLE and BE of (β_1, β_2, β_3) are very close to the normal distribution. Remember that the skewness and kurtosis of the normal distribution are given by zero and three, respectively.

As for σ^2, AVE of BE is closer to the true value than that of MLE, because AVE of MLE is 0.752 (see Table 3) and that of BE is 1.051 (see Table 2). However, in the SER and RMSE criteria, MLE is superior to BE, since SER and RMSE of MLE are given by 0.276 and 0.372 (see Table 3) while those of BE are 0.380 and 0.384 (see Table 2). The empirical distribution obtained from 10^4 estimates of σ^2 is skewed to the right (Skewness is positive for both MLE and BE) and has a larger kurtosis than the normal distribution because Kurtosis is greater than three for both tables.

For ρ, AVE of MLE is 0.559 (Table 3) and that of BE is given by 0.661 (Table 2). As it is also seen in Figures 3 and 1, BE is less biased than MLE from the AVE criterion. Moreover, SER and RMSE of MLE are 0.240 and

Parameter	β_1	β_2	β_3	ρ	σ^2
True Value	10	1	1	0.9	1
AVE	10.012	0.999	1.000	0.559	0.752
SER	3.025	0.171	0.053	0.240	0.276
RMSE	3.025	0.171	0.053	0.417	0.372
Skewness	0.034	−0.045	−0.008	−1.002	0.736
Kurtosis	2.979	3.093	3.046	4.013	3.812
5%	5.096	0.718	0.914	0.095	0.363
10%	6.120	0.785	0.933	0.227	0.426
25%	7.935	0.883	0.965	0.426	0.550
50%	10.004	0.999	1.001	0.604	0.723
75%	12.051	1.115	1.036	0.740	0.913
90%	13.913	1.217	1.068	0.825	1.120
95%	15.036	1.274	1.087	0.863	1.255

Table 2: MLE: $n = 20$ and $\rho = 0.9$

Parameter	β_1	β_2	β_3	ρ	σ^2
True Value	10	1	1	0.9	1
AVE	10.010	0.999	1.000	0.661	1.051
SER	2.782	0.160	0.051	0.188	0.380
RMSE	2.782	0.160	0.051	0.304	0.384
Skewness	0.008	−0.029	−0.022	−1.389	0.725
Kurtosis	3.018	3.049	2.942	5.391	3.783
5%	5.498	0.736	0.915	0.285	0.515
10%	6.411	0.798	0.934	0.405	0.601
25%	8.108	0.891	0.966	0.572	0.776
50%	10.018	1.000	1.001	0.707	1.011
75%	11.888	1.107	1.036	0.799	1.275
90%	13.578	1.205	1.067	0.852	1.555
95%	14.588	1.258	1.085	0.875	1.750

Table 3: BE with $M = 5000$ and $N = 10^4$: $n = 20$ and $\rho = 0.9$

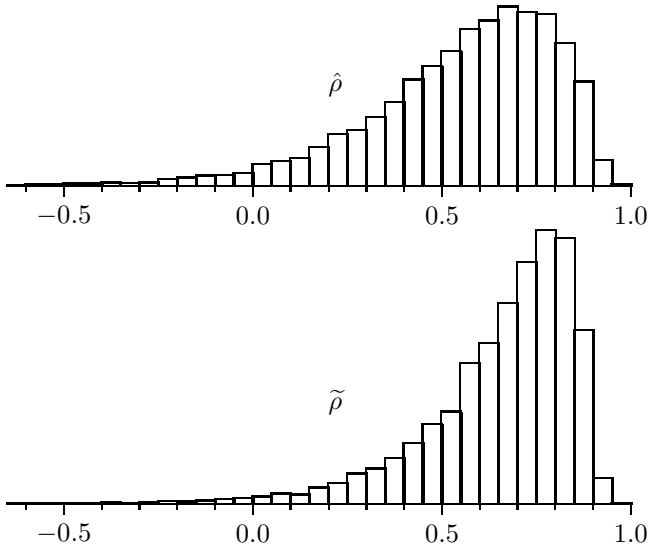

Figure 3: Empirical distributions for MLE and BE
—— $n = 20$, $\rho = 0.9$, $M = 5000$, $N = 10^4$ and $L = 10^4$ ——

0.417, while those of BE are 0.188 and 0.304. Therefore, BE is more efficient than MLE. Thus, in the AVE, SER and RMSE criteria, BE is superior to MLE with respect to ρ. The empirical distributions of MLE and BE of ρ are skewed to the left because Skewness is negative, which value is given by -1.002 in Table 3 and -1.389 in Table 2. We can see that MLE is less skewed than BE. For Kurtosis, both MLE and BE of ρ are greater than three and therefore the empirical distributions of the estimates of ρ have fat tails, compared with the normal distribution. Since Kurtosis in Table 2 is 5.391 and that in Table 3 is 4.013, the empirical distribution of BE has more kurtosis than that of MLE. Figure 3 also indicates these facts.

Figure 3 corresponds to $\hat{\rho}$ in Table 3 and $\widetilde{\rho}$ in Table 2, respectively. As we can see from Skewness and Kurtosis in Table 3 and Table 2, $\hat{\beta}_i$ and $\widetilde{\beta}_i$, $i = 1, 2, 3$, are very similar to a normal distribution. From Skewness and Kurtosis, $\hat{\sigma}^2$ and $\widetilde{\sigma}^2$ are quite different from a normal distribution, but they are very similar to each other. Also, $\hat{\rho}$ and $\widetilde{\rho}^2$ are quite different from the normal distribution. However, the empirical distribution of $\hat{\rho}$ is quite different from that of $\widetilde{\rho}$. We can observe that $\widetilde{\rho}$ is more skewed to the left than $\hat{\rho}$ and $\widetilde{\rho}$ has a larger kurtosis than $\hat{\rho}$. Therefore, the empirical distributions of $\hat{\rho}$ and $\widetilde{\rho}$ are shown in Figure 3. The mode of MLE $\hat{\rho}$ lies on the interval between 0.65 and 0.70, while that of BE $\widetilde{\rho}$ is between 0.75 and 0.80. From the facts that SER of $\widetilde{\rho}$ is smaller than

that of $\hat{\rho}$ and that the mode of $\widetilde{\rho}$ is closer to the true value than that of $\hat{\rho}$, $\widetilde{\rho}$ is distributed around the true value 0.9.

4. Conclusion

In this paper, we have compared MLE with BE, using the regression model with the autocorrelated error term. Chib [1] and Chib and Greenberg [2] applied the Gibbs sampler to the autocorrelation model, where the initial density of the error term is ignored. Under this setup, the posterior distribution of ρ reduces to the normal distribution. Therefore, random draws of ρ given β, σ^2 and (y_t, X_t) can be easily generated. However, when the initial density of the error term is properly taken into account, the posterior distribution of ρ is not normal and it cannot be represented in an explicit functional form. Accordingly, in this paper, the MH algorithm has been applied to generate random draws of ρ from its posterior density.

The obtained results are summarized as follows. Given $\beta' = (10, 1, 1)$ and $\sigma^2 = 1$, in Figure 3 we have the relationship between ρ and $\hat{\rho}$, and $\widetilde{\rho}$ corresponding to ρ is drawn in Figure 1. In the two figures, we can observe: (i) both MLE and BE approach the true parameter value as n is large, and (ii) BE is closer to the 45° degree line than MLE and accordingly BE is superior to MLE.

Moreover, we have compared MLE with BE in Table 3 and Table 2, where $\beta' = (10, 1, 1)$, $\rho = 0.9$ and $\sigma^2 = 1$ are taken as the true values. As for the regression coefficient β, both MLE and BE gives us the unbiased estimators. However, we have obtained the result that BE of β is more efficient than MLE. For estimation of σ^2, BE is less biased than MLE. In addition, BE of the autocorrelation coefficient ρ is also less biased than MLE. Therefore, as for inference on β, BE is superior to MLE, because it is plausible to consider that the estimated variance of $\hat{\beta}$ is biased much more than that of $\widetilde{\beta}$. Remember that variance of $\hat{\beta}$ depends on both ρ and σ^2. Thus, from the simulation studies, we can conclude that BE performs much better than MLE.

References

[1] S. Chib, Bayes regression with autoregressive errors: A Gibbs sampling approach, *Journal of Econometrics*, **58** (1993), 275-294.

[2] S. Chib, E. Greenberg, Bayes inference in regression models with ARMA(p, q) Errors, *Journal of Econometrics*, **64** (1994), 183-206.

[3] D.W.K. Andrews, Exactly median-unbiased estimation of first order autoregressive unit root models, *Econometrica*, **61** (1993), 139-165.

[4] H. Tanizaki, Bias correction of OLSE in the regression model with lagged dependent variables, *Computational Statistics and Data Analysis*, **34** (2000), 495-511.

[5] H. Tanizaki, On least-squares bias in the AR(p) models: Bias correction using the bootstrap methods, Unpublished Manuscript (2001), http://ht.econ.kobe-u.ac.jp/ tanizaki/cv/working/unbiased.pdf

[6] J.M. Bernardo, A.F.M. Smith, *Bayesian Theory*, John Wiley & Sons (1994).

[7] B.P. Carlin, T.A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall (1996).

[8] M.H. Chen, Q.M. Shao, J.G. Ibrahim, *Monte Carlo Methods in Bayesian Computation*, Springer-Verlag (2000).

[9] D. Gamerman, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman & Hall (1997).

[10] C.P. Robert, G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag (1999).

[11] A.F.M. Smith, G.O. Roberts, Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society*, Ser. B, **55** (1993), 3-23.

[12] G. Judge, C. Hill, W. Griffiths, W. Lee, *The Theory and Practice of Econometrics*, John Wiley & Sons (1980).

[13] J. Geweke, Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, In: *Bayesian Statistics* (Ed-s: J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith), Oxford University Press, Vol. **4**, 169-193 (with discussion) (1992).

[14] K.L. Mengersen, C.P. Robert, C. Guihenneuc-Jouyaux, MCMC Convergence Diagnostics: A Reviewww, In: *Bayesian Statistics* (Ed-s: J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith), Oxford University Press, Vol. **6**, 514-440 (with discussion) (1999).

Parameter	β_1	β_2	β_3	ρ	σ^2
True Value	10	1	1	0.9	1
(a) $(M, N) = (5000, 5000)$					
AVE	10.011	0.999	1.000	0.661	1.051
RMSE	2.785	0.160	0.052	0.305	0.384
50%	10.015	1.000	1.001	0.707	1.011
Skewness	0.004	−0.027	−0.022	−1.390	0.723
Kurtosis	3.028	3.056	2.938	5.403	3.776
(b) $(M, N) = (1000, 10^4)$					
AVE	10.010	0.999	1.000	0.661	1.051
RMSE	2.783	0.160	0.051	0.304	0.384
50%	10.014	1.000	1.001	0.706	1.011
Skewness	0.008	−0.029	−0.021	−1.391	0.723
Kurtosis	3.031	3.055	2.938	5.404	3.774

Table 4: BE with $(M, N) = (5000, 5000)$, $(1000, 10^4)$: $n = 20$ and $\rho = 0.9$

5. Appendix: Is $(M, N) = (5000, 10^4)$ Sufficiently Large?

In this appendix, we check whether M and N are enough large in BE. Comparison between Table 2 and Table 4(a) shows whether $N = 5000$ is large enough and we can see from Table 2 and Table 4(b) if the burn-in period $M = 1000$ is large enough. For the burn-in period M, there are some diagnostic tests, which are discussed in Geweke [13] and Mengersen, Robert and Guihenneuc-Jouyaux [14]. However, since their tests are applicable in the case of one sample path, we cannot utilize them in this section. Because L simulation runs are implemented in this paper, we have L test statistics if we apply the tests. It is not possible to evaluate L testing results at the same time. Therefore, we consider using the alternative approach to see if $M = 1000$ and $N = 5000$ are sufficient. We can conclude that $N = 5000$ is large enough if Table 2 is very close to Table 4(a) and that $M = 1000$ is enough if Table 2 is close to Table 4(b).

From Table 2 and Table 4, we can observe as follows. The difference between Table 2 and Table 4(a) is at most 0.012 (see Kurtosis in ρ) and that between Table 2 and Table 4(b) is less than or equal to 0.013 (see Kurtosis in β_1 and ρ). Thus, all the values in the tables are very close to each other. Therefore, we can conclude that $(M, N) = (1000, 5000)$ is enough. For safety, we take the case of $(M, N) = (5000, 10^4)$ in this paper.