BAYESIAN ANALYSIS OF THE SIMPLE LINEAR REGRESSION WITH MEASUREMENT ERRORS

Marta Yukie BABA[1]
Fernando Antonio MOALA[1]

- *ABSTRACT: Usually the classical approach to make inference in linear regression model assumes that the independent variable does not contain measurement errors. In practice, however, the data can contain measurement errors and the presence of these errors can affect the results of the analysis drastically. Rodrigues and Baba (1994) proposed a Bayesian approach to estimate the slope parameter β in linear regression model with measurement errors considering the reliability ratio K_X as known. There are situations, however, where the information regarding the reliability ratio K_X not always is available. In this paper, our main interest is to make a Bayesian inference about β under the assumption that the reliability ratio K_X is unknown. To obtain the posterior distribution we use Gibbs Sampler algorithm.*

- *KEYWORDS: Posterior distribution; reliability ratio; slope parameter; Gibbs sampler.*

1 Introduction

The classical simple linear regression analysis assumes that the independent variable is defined by

$$y_i = \alpha + \beta x_i + \varepsilon_i, i=1, ...,n; \quad (1)$$

where $(x_1, \ldots, x_n)$ is fixed in repeated sampling and ε_i are independent $N(0, \sigma^2_\varepsilon)$ random variables. It is assumed that x_i is measured without error. However, in practice, particularly in social sciences and biological essay, this assumption is often violated. There is a lot of work on the problem of parameter estimation when the x_i contain errors of measurement, see for example, Fuller (1987). In the present paper we propose a Bayesin approach to estimate the model parameters. We shall study models of type (1), with $\alpha = 0$, where instead of observing x_i one observes the sum

$$X_i = x_i + u_i, i =1, ...,n. \quad (2)$$

We make the assumption that

[1]UNESP, Department of Mathematics, Presidente Prudente, SP, Brazil. E-mail: *marta@fct.unesp.br / femoala@fct.unesp.br*

$$(x_i, u_i, \varepsilon_i)' \sim NI((\mu_x, 0, 0)', diag(\sigma_x^2, \sigma_u^2, \sigma_\varepsilon^2)); \tag{3}$$

where $\sim NI$ is an abbreviation for "distributed normally and independently", and $diag\left(\sigma_x^2, \sigma_e^2, \sigma_u^2\right)$ is a diagonal matrix with the given elements on the diagonal.

It follows from (1), (2) and (3), that the vector $(X_i\ ,\ Y_i)'$ is distributed as a bivariate normal

$$\begin{bmatrix} Y_i \\ X_i \end{bmatrix} \sim NI\left(\begin{bmatrix} \beta\mu_x \\ \mu_x \end{bmatrix}, \begin{bmatrix} \beta^2\sigma_x^2 + \sigma_e^2 & \beta\sigma_x^2 \\ \beta\sigma_x^2 & \sigma_x^2 + \sigma_u^2 \end{bmatrix}\right). \tag{4}$$

We can without loss of generality take $\mu_x = 0$.

A way of inference of the parameter consists of analyzing it subject to the two-dimensional model, however, Rodrigues and Cordani (1990) analyze under other perspective, working with the conditional distribution of $Y_i$given X_i. Because $(X_i\ ,\ Y_i)$ is distributed as a bivariate normal, the conditional distribution of Y_i given X_i is given by

$$Y_i \Big| X_i \quad \sim \quad N\left[K_X\, \beta\, X_i\ ,\ K_X\, \beta^2\sigma_u^2 + \sigma_e^2\right]; \tag{5}$$

where $K_X = \dfrac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \dfrac{\sigma_x^2}{\sigma_X^2}$ is called reliability ratio.

Rodrigues and Baba (1994) proposed a Bayesian approach to estimate the slope parameter β in linear regression model with measurement errors considering the reliability ratio K_X as known.

There are situations, however, where the information regarding the reliability ratio K_X not always is available. In this paper, our main interest is to make a Bayesian inference about β under the assumption that the reliability ratio K_X is unknown.

2. Bayesian analysis of the regression model

Let $(X_1\ ,\ Y_1)$, . . ., $(X_n\ ,\ Y_n)$ are independent and identically distributed random variables in agreement with the conditional model:

$$Y_i \Big| X_i \quad \sim \quad N\left[K_X\, \beta\, X_i\ ,\ K_X\, \beta^2\sigma_u^2 + \sigma_e^2\right];$$

where K_X is unknown.

We consider the "over identifiable" situation, that is, $\sigma_e^2 = \sigma_u^2 = \sigma_0^2$, where the common variance is known. Thus, the likelihood function is given by

$$L(\beta, K_X) \propto \left(K_X\, \beta^2 + 1\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_0^2\left(K_X\, \beta^2 + 1\right)}\sum_{i=1}^{n}\left(Y_i - \beta\, K_X X_i\right)^2\right\}. \tag{6}$$

Denoting the variance of the distribution given in (6) by $\sigma^2 = K_X \beta^2 \sigma_u^2 + \sigma_e^2 = \sigma_0^2\left(K_X \beta^2 + 1\right)$, we obtain

$$L(\beta, K_X, \sigma^2) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta K_X X_i\right)^2\right\}. \tag{7}$$

Box and Tiao (1973) propose a locally uniform joint prior distribution for β and $\log \sigma$ given by:

$$\pi\left(\beta,\ \log\sigma^2\right) \propto \text{ constant.} \tag{8}$$

Consequently, the joint prior density for β and σ^2 is:

$$\pi\left(\beta, \sigma^2\right) \propto \frac{1}{\sigma^2}. \tag{9}$$

To propose a prior distribution for K_X, we observe that a close value of zero can be due to a great measurement error $\left(\sigma_X^2 = \infty\right)$, an error of planning of the data that a statistician cannot accept, or it can be due to $\sigma_x^2 = 0$ established for the functional case, a situation no studied here.

Thus, we consider a prior uniform in (0.3, 1) for K_X, that is,

$$\pi\left(K_X\right) \propto \text{ constant}, 0.3 < K_X < 1. \tag{10}$$

Now, supposing the parameters are independent we obtain a joint prior for $\left(\beta, K_X, \sigma^2\right)$ given by

$$\pi\left(\beta, K_X, \sigma^2\right) \propto \frac{1}{\sigma^2}. \tag{11}$$

Therefore, from (7) and (11) we can express the joint posterior distribution as:

$$p\left(\beta, K_X, \sigma^2 \mid \text{data}\right) \propto \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta K_X X_i\right)^2\right\}. \tag{12}$$

Because our intention is to estimate the slope parameter, we consider the application of Gibbs Sampler to obtain the marginal posterior distribution of β. For this, we need to obtain the complete conditional distributions.

Writing $\sum_{i=1}^{n}\left(Y_i - \beta K_X X_i\right)^2 = \sum_{i=1}^{n} Y_i^{\,2} - 2\beta K_X \sum_{i=1}^{n} X_i Y_i + \sum_{i=1}^{n} Y_i^{\,2}$ and after some algebras, the posterior conditional distribution $p\left(\beta \mid K_X, \sigma^2, \text{data}\right)$ is given by

$$p\left(\beta \middle| K_X, \sigma^2, \text{data}\right) \propto \exp\left\{-\frac{1}{2\sigma^2} K_X^2 \sum_{i=1}^{n} X_i^2 \left(\beta - \frac{\sum_{i=1}^{n} X_i Y_i}{K_X \sum_{i=1}^{n} X_i^2}\right)^2\right\}; \quad (13)$$

that is, $\beta \middle| K_X, \sigma^2, \text{data} \sim N\left(\frac{\sum_{i=1}^{n} X_i Y_i}{K_X \sum_{i=1}^{n} X_i^2}; \frac{\sigma^2}{K_X^2 \sum_{i=1}^{n} X_i^2}\right)$.

The posterior conditional distribution for K_X given by

$$p\left(K_X \middle| \beta, \sigma^2, \text{data}\right) \propto \exp\left\{-\frac{1}{2\sigma^2} \beta^2 \sum_{i=1}^{n} X_i^2 \left(K_X - \frac{\sum_{i=1}^{n} X_i Y_i}{\beta \sum_{i=1}^{n} X_i^2}\right)^2\right\}, \; 0.3 \le K_X \le 1; \quad (14)$$

which is a Normal Truncated with K_X restricted to the interval (0.3, 1).

Besides, conditional to β and K_X, the posterior conditional distribution for σ^2 is given by

$$p\left(\sigma^2 \middle| \beta, K_X, \text{data}\right) \propto \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(Y_i - \beta K_X X_i\right)^2\right\}; \quad (15)$$

that is, the posterior conditional distribution for σ^2 is a Gamma Inverted distribution with parameters $\alpha = \frac{n}{2}$ and $\gamma = \frac{\sum_{i=1}^{n} \left(Y_i - \beta K_X X_i\right)^2}{2}$.

3. Implementation of Gibbs Sampler

Gibbs sampler is a particular case of substitution sampling Gelfand and Smith (1990) in which all full conditional densities are supposed known. A nice introduction to the Gibbs sampler is given by Casella and George (1992). In this paper the conditional densities are given by equations (13), (14) and (15).

We now describe the Gibbs Sampler implementation used in our framework. The algorithm proceeds as follows

1. Choose starting values $(\beta^0, K_X^0, \sigma_0^2)$.

2. At step i+1:

a) Draw β^i from the conditional posterior $p\left(\beta \middle| K_X^i, \sigma_i^2, \text{data}\right)$ given in (13);

b) Draw K_X^i from the conditional posterior $p\left(K_X \left| \beta^{i+1}, \sigma_i^2, \text{data}\right.\right)$ given in (14);

c) Draw σ_i^2 from the conditional posterior $p\left(\sigma^2 \left| \beta^{i+1}, K_X^{i+1}, \text{data}\right.\right)$ given in (15).

3. This provides a sequence of sampled values $(\beta^i, K_X^i, \sigma_i^2)$, i=0, 1, . . ., N which is a realization of the Markov chain associated with the full conditional densities of $p\left(\beta, K_X, \sigma^2 \left| \text{data}\right.\right)$ given in (12).

4. Numerical illustration

In this section, we illustrated the performance of the procedure proposed in this paper based in the samples generated in the software R considering α = 0, $\beta = 2$ and variances $\sigma_e^2 = \sigma_u^2 = \sigma_0^2 = 1$ and $\sigma_x^2 = 1$. Two values for the reliability ratio K_X are considered, for instance K_X = *0.5* and K_X =0.8, in order to compare the performance of estimates in the analysis. In this case, the parameter σ assumes values equal toσ = 1.732 and 2.05, respectively. A practical example with real data is also presented.

As we are not able to find an analytic expression for marginal posterior distributions and hence to extract characteristics of parameters such as Bayes estimates, and credible intervals, we need to appeal to the Gibbs Sampler algorithm to obtain a sample of values of parameters from the joint posterior. The chain is run for *N*=10000 iterations with a burn-in period of size 1000, which were discarded to eliminate the effect of the initial values.

Figure 1 presents the MCMC output plot and the marginal densities resulting for the parameters of the regression model considering K_X = *0.5* and *n =10*. The MCMC plots suggest we have achieved convergence. The posterior summaries of interest are given in Tables1 and2 for different sample sizes as *n=10* and *n=50*,respectively.

Both Tables allow a comparison of the estimators of β, K_X and σ using the approaches proposed by Rodrigues and Cordani (1990) and the Bayesian procedure proposed in this paper. The 95% intervals from the maximum likelihood (ML) and Bayesian approaches are also displayed in the Table 1.

The results from Table 1 show that ML and Bayesian approaches do not provide good estimates for β, however the confidence interval by ML approach has negative values. The Bayesian estimate for the parameter K_X is too close to the true value while the ML estimate assumes an impossible value, that is, greater than 1. Besides, the ML interval is larger than the range (0, 1) of K_X. For the parameter σ, the ML estimate is better than the Bayesian.

Table 1 - Estimates values and confidence intervals for the parameters β, K_X and σ by ML and Bayesian approaches for = *0.5* and *n =10*

	β	K_X	σ
ML	1.3254 (-0.5703, 3.2211)	1.1013 (-0.76736, 2.9699)	1.7131 (0.4504, 2.9757)
Bayesian	2.8087 (0.8226, 5.8844)	0.5842 (0.3105, 0.9623)	1.8658 (1.1784, 3.1757)

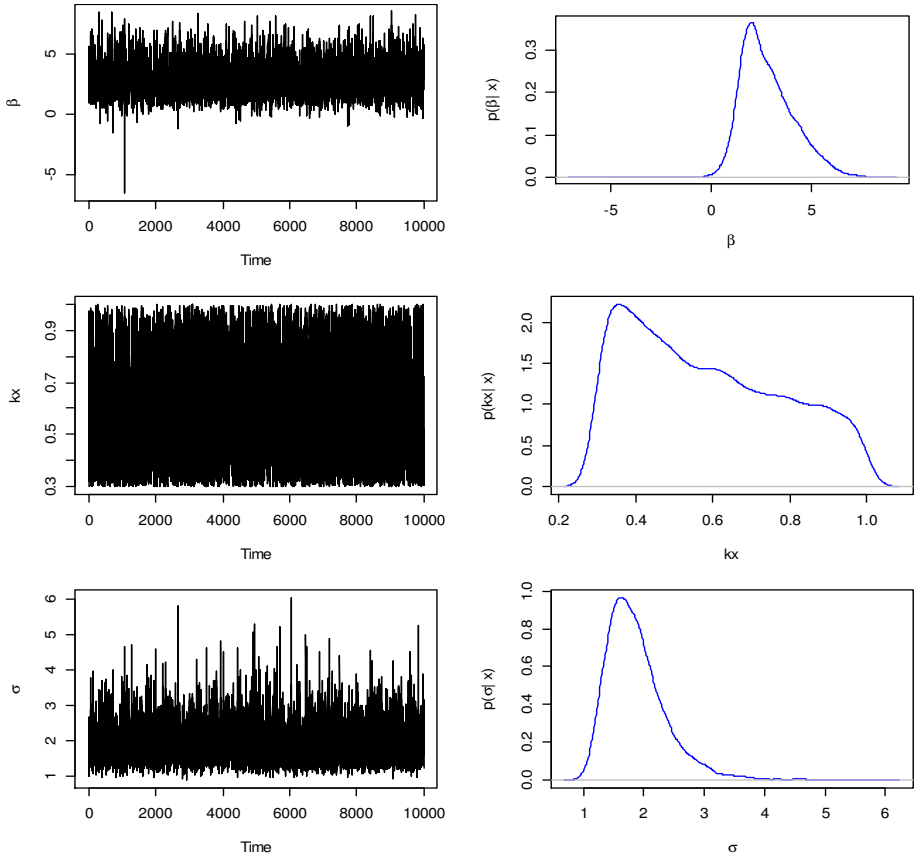

Figure 1 - Estimates Densities via Gibbs Sampler and the traces of the chains generated for the parameters β, K_X and σ.

Table 2 - Estimates values and confidence intervals of the parameters β, K_X and σ by ML approach and Bayesian estimator obtained via Gibbs Sampler for $K_X = 0.5$ and $n=50$

	β	K_X	σ
ML	1.6666 (0.5069, 2.8263)	0.5785 (0.08641, 1.0707)	1.6146 (1.0824, 2.1469)
Bayesian	1.8237 (0.8717, 3.4150)	0.5742 (0.3085, 0.9729)	1.6405 (1.3478, 2.0044)

When the sample size n increases ($n=50$) all the estimates get better and more accurate providing quite similar results, as expected. Note however that the ML estimate for K_X becomes close to the 0.5 but with interval still large while the Bayesian estimation for K_X does not improve.

In summary, when sample size n is small, the ML estimates can assume inadmissible values and the Bayesian estimates will have likely values but not very accurate. However, for moderate n ($n = 50$), the most estimates will have similar values.

In Tables 3 and 4 we can analyze the point and intervals estimates of β, K_X and σ for different values of K_X , (K_X=0.5 and K_X=0.8) for a sample size n=10. We consider $K_X = 0.8$ resulted from $\sigma_e^2 = \sigma_u^2 = 1$ and $\sigma_x^2 = 4$.

Table 3 -Estimates values of the parameters β, K_X and σ by ML approach and Bayesian estimator (via Gibbs Sampler) for n =10

	β	K_X	σ
ML	0.8252 (1.3254)	2.3598 (1.1013)	1.6146 (1.7131)
Gibbs	3.7268 (2.8087)	0.5738 (0.5842)	1.7682 (1.8658)

Note: values corresponding to K_X =0.5 (are between parentheses).

Table 4 - 95% confidence intervals of the parametersβ, K_X and σ by ML approach and Bayesian estimator (via Gibbs Sampler) for n =10

	β	K_X	σ
ML	(-0.3606, 2.0110)	(-1.1349, 5.8545)	(0.4245, 2.8047)
Gibbs	(1.8296, 6.6435)	(0.3099, 0.9662)	(1.1069, 2.9527)

As the sample size *n* is small, the ML and Bayesian estimates have the similar behavior for K_X=*05*, however, by comparing the estimations for K_X=*0.5* and K_X=*0.8* we analyze how the parameter K_X can affect the estimations. Note that for K_X=0.8 ($\sigma_x^2 = 4$) the estimation of β gets worse and inaccurate with both estimation approaches. For parameter K_X the Bayesian approach is still better than ML.

Therefore K_X is an important parameter of interest in the regression analysis with measure errors and it should be considered in this study, mainly when there are few observed data set.

5. A real illustration with literature data

Consider the example of regression model with measurement errors proposed by Fuller (1987, page 18) that involves yield of corn (Y) for different levels of soil nitrogen (X). Here the explanatory variable, soil nitrogen level, has been determined with measurement error.

It is assumed there is a prior estimate of the measurement error for soil nitrogen is $\sigma_u^2 = 57$. We consider the "over identifiable" situation, that is, $\sigma_e^2 = \sigma_u^2 = \sigma_0^2$.

The data set given in Table 5 represents the corn and determinations available soil nitrogen collected at 11 sites on Marshall Soil in Iowa.

Table 5 -Yields of corn on Marshall soil in Iowa

Site	1	2	3	4	5	6	7	8	9	10	11
Yield (Y)	86	115	90	86	110	91	99	96	99	104	96
Soil Nitrogen (X)	70	97	53	64	95	64	50	70	94	69	51

After run the Gibbs Sampler algorithm for N=30000 iterations we provide the Figure 2 with the MCMC output plot and the marginal densities resulting for the parameters of the regression model.

Comparison of the regression estimates from ML and Bayesian approaches are provided at the Table 6 below which shows the point estimates and 95% confidence intervals for the parametersβ, K_X and σ.

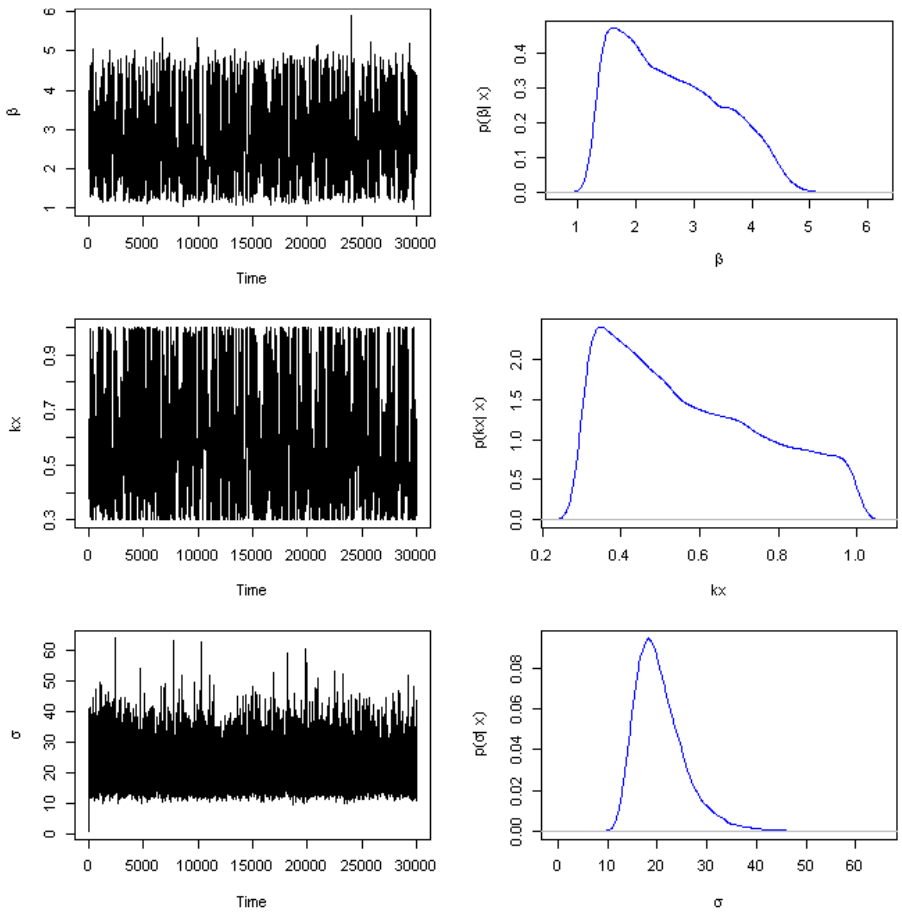

Figure 2 - Estimates Densities via Gibbs Sampler and the traces of the chains generated for the parametersβ, K_X and σ.

Table 6 - Estimates values and confidence intervals of the parameters β, K_X and σ by ML approach and Bayesian estimator obtained via Gibbs Sampler for *n=11*

	β	K_X	σ
ML	4.0292	0.3289	19.0091
	(3.5534, 4.5051)	(0.2518, 0.4059)	(5.6499, 32.3683)
Bayesian	2.5170	0.5970	20.6173
	(1.3194, 4.3654)	(0.3098, 0.9734)	(13.2046, 33.5342)

Now, what would happen if we did not consider the estimation error in the variable X. In other words, what would be the estimator of β for X fixed in a simple linear regression model? We can specify such a linear regression model easily by R software and estimation results would be:

Call:
lm(formula = Y ~ -1 + X)

Residuals:
Min	Median	3Q	Max
-25.566	-10.152	3.238	16.165	32.742

Estimate	Std. Error	t	value Pr(>|t|)
X 1.32517	0.07898	16.78	1.19e-08

IC 95% (1.278496 ; 1.371844)

Conclusion

Because the procedures proposed by Rodrigues and Cordani (1990) are based in asymptotic results, the classic estimator of β in the regression model with measurement errors does not produce satisfactory estimates when the size of the sample is small. In this case, the Bayesian approach for estimation of the studied model produces better results than classical estimators. Therefore, we verified that Bayesian method usually requires less sample data to achieve the better quality of inferences than the method based on classic theory. In many cases, this is the practical motivation for using Bayesian methods and represents the practical advantage in the use of prior information. This is an especially important consideration in those areas of application where sample data may be either expensive or difficult to obtain it.

In addition, the statistical inferences based on sampling theory are usually more restrictive than Bayesian Inference due to the exclusive use of sample data. The Bayesian Inference's use of relevant past experience, which is quantified by the prior distribution, produces more informative inferences in those cases where the prior distribution accurately reflects the variation in the parameter. So, it was possible that we put the information that K_X assumes values between 0 and 1 in this work. We observe that K_X is an important parameter in the regression analysis with measure errors and it should be considered in the study, mainly when there are few observed data set. The degree to which more informative inferences occur otherwise depends upon the quality of the assessments embodied in the prior distribution. Therefore, other priors could be tried to improve the estimation. The comparison of priors for the model with errors in variables is a future study of our interest.

BABA, M. Y.; MOALA, F. A. Análise Bayesiana da regressão linear simples com erros de medida. *Rev. Bras. Biom.*, São Paulo, v.30, n.2, p.174-184, 2012.

- RESUMO:Geralmente a análise clássica de inferência do modelo de regressão linear assume que a variável independente não contém erros de medida.Na prática, porém, os dados podem conter erros de medição e a presença destes erros pode afetar drasticamente os resultados da análise. Rodrigues e Baba (1994) propuseram uma abordagem Bayesiana para estimar o parâmetro de inclinação β no modelo de regressão linear com erros de medida, considerando a razão de confiabilidade K_X como conhecida. Há situações, no entanto, em que a informação sobre a razão de confiabilidade K_X nem sempre é disponível. Neste artigo,nosso interesse principal é realizar uma inferência Bayesiana do parâmetro β sob a suposição deque a razão de confiabilidade K_X é desconhecida. Para obter a distribuição a posteriori usamos o algoritmo amostrador de Gibbs.
- PALAVRAS-CHAVE:Distribuição a posteriori; razão de confiabilidade; parâmetro de inclinação; amostrador de Gibbs.

References

BABA, M. Y. *Inferência bayesiana para regressão linear simples com erro nas variáveis*.1994. Dissertação (Mestrado) – Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, São Carlos, 1994.

CASELLA, G.; GEORGE, E. I. Explaining the Gibbs sampler. *Am. Stat.*, Baltimore, v.46, n.3, p.167-174, 1992.

FULLER. W.A. *Measurement error models*. New York: J. Wiley, 1987. 439p.

GELFAND, A.E.; SMITH, A.F.M. Sampling based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, Baltimore, v.85, n.4, p.398-410, 1990.

RODRIGUES, J.; BABA, M.Y. Bayesian estimation of a simple regression model with measurement error. *Braz. J. Prob. Stat.*, São Paulo, v.8, n.2, p.107-118, 1994.

RODRIGUES, J.; CORDANI, L.K.A note on likelihood estimation of a simple regression model with measurement error, via the orthogonal parameterization. *S. Afr. Stat. J.*, Pretoria, n.24, p.177-83, 1990.

Received in 26.03.2012

Approved after revised in 08.08.2012