Bayesian Estimation Analysis of Bernoulli Measurement Error Model for Longitudinal Data

Dewang Li, Meilan Qiu*, Zhongyi Ke
School of Mathematics and Statistics, Huizhou University, Huizhou, Guangdong, 516007, China.

* Corresponding author. Tel: 86-18319620962; email: qml_1981@126.com
Manuscript submitted March 2, 2020; accepted May 28, 2020.
doi: 10.17706/ijapm.2020.10.4.160-166

Abstract: The Bayesian method is used to study the inference of the semi-parametric measurement error model (MEs) with longitudinal data. A semi-parametric Bayesian method combined with fracture prior and Gibbs sampling combined with Metropolis-Hastings (MH) algorithm is applied and applied to the simulation observation from the posterior distribution, and the combined Bayesian statistics of unknown parameters and measurement errors are obtained. We obtained Bayesian estimates of the parameters and covariates of the measurement error model. Under three different priori assumptions, four simulation studies illustrate the effectiveness and utility of the proposed method.

Key words: Bayesian, measurement error, longitudinal data, Bernoulli model.

1. Introduction

Longitudinal data is obtained when the same individual is repeatedly measured at different time points. Longitudinal data is widely found in bio-medicine, epidemiology and labor medicine. For example, biomedical longitudinal samples can generally be obtained through clinical trials and observational cohort studies. Longitudinal data is also widely used in the fields of finance, economy, etc. It is an unbalanced data and is generally processed using a linear hybrid model. Measurement error data and missing data are often encountered for various reasons. When the covariate contains measurement error, You (2006) [1] proposed a profile least squares estimation method for error correction; Zhou (2009) [2] the covariate of measurement error in the research model has statistical inference problem of auxiliary information; Wei (2010) [3] studied the parameter estimation problem of the model when the response variable is missing and the covariate contains the measurement error; Wei (2012) [4] studied the constraint estimation and hypothesis testing of the variable coefficient partial linear measurement error model parameters. There are also Liang, Härdel and Carroll (1999) [5], Ma and Carroll (2006) [6], Liang, Wang and Carroll (2007) [7], Pan, Zeng and Lin (2008) [8] and other literature pairs. Such models have been studied. This paper proposes a hybrid algorithm for generating the observations required for Bayesian inference from the parameter posterior distribution and from the covariates of the ME. The algorithm combines a normal distribution with a mixed normal distribution. Gibbs sampling of a priori and MH algorithms was broken.

2. The Measurement Error Model

For $i = 1, \ldots, n$, hypothesis Y_i is the observation variable, which X_i is an unobservable covariate vector of order$r \times 1$, and U_i is a covariate vector that can be observed in one order $p \times 1$. Let $Z_i = (X_i^T, U_i^T)^T$, we

assume that the values Y_i are conditionally independent of each other. For longitudinal data, we consider the following generalized linear measurement error model of structure.

$$P(Y_i|X_i,\psi) = \boldsymbol{exp}\{\frac{Y_i\tau_i - d(\tau_i)}{\psi} + c(Y_i,\psi)\}. \quad (1)$$

Here $\gamma_i = E(Y_i|X_i) = \dot{d}(\tau_i)$, ψ is a divergence parameter, $d(\cdot)$ and $c(\cdot,\cdot)$ are specific differentiable functions, and has $\dot{d}(\alpha_i) = \partial d(\tau_i)/\partial\tau_i$ and $\ddot{d}(\tau_i) = \partial^2 d(\tau_i)/\partial\tau_i^2$. The conditional mean γ_i satisfies the following equation.

$$\lambda_i = g(\gamma_i) = X_i^T\beta_z + U_i^T\beta_v = Z_i^T\beta. \quad (2)$$

Here $g(\cdot)$ is a monotonic differentiable link function, which $\beta = (\beta_z^T, \beta_v^T)^T$ is an unknown regression coefficient with vector $(r+p)\times 1$. According to reference [9], for each individual *i*, we measure m times for the true value covariate X_i. Y_i and X_i are error independent. That is, for each $j = 1,\dots,m$ with following equations, we can't observe $X_i$but we can observe W_{ij}.

$$W_{\text{ij}} = X_i + \eta_{ij}. \quad (3)$$

These measurement error values $\eta_{\text{ij}}$are subject to unknown distributions, and they are independent of the true values X_i. According to Lachos (2010) [10], our hypothetical distribution η_{ij} is suitable for a mixed model of the Dirichlet Process (DP).

In order to calculate the previously set covariate measurement error model, we also need to define a real covariate model. The true covariate model for X_{ki} $(k = 1,\dots,r)$ can be defined as

$$X_{\text{ki}} = \alpha_{k0} + \alpha_{kv}^T U_i + \xi_{ki},\ \xi_{ki}\sim N(0,\sigma_z^2). \quad (4)$$

Here α_{k0} is an intercept, which$\alpha_{kv} = (\alpha_{k1},\dots,\alpha_{kp})^T$ is an order of $p\times 1$unknown regression coefficient vectors. Let $Y = \{Y_1,\dots,Y_n\}$, $X = \{X_1,\dots,X_n\}$, $U = \{U_1,\dots,U_n\}$, $\eta = \{\eta_1,\dots,\eta_n\}$ and $W = \{W_1,\dots,W_n\}$, $X_i = (X_{1i},\dots,X_{ri})^T$, $\eta_i = (\eta_{i1},\dots,\eta_{im})$ and $W_i = \{W_{i1},\dots,W_{im}\}$, for each$i = 1,\dots,n$.Suppose$\varepsilon_y = \{\beta,\psi\}$, $\varepsilon_\alpha = \{\alpha_{10},\dots,\alpha_{r0},\alpha_{1v},\dots\alpha_{rv},\sigma_z^2\}$ and$\varepsilon = \{\varepsilon_y,\varepsilon_\alpha,\varepsilon_\eta\}$, $\varepsilon_\eta$is the parameter of equation (3). The joint probability density function for$\{Y,W,\theta,X\}$ representation

$$P(Y,W,\theta,X|U,\varepsilon) = \prod_{i=1}^n\{P(Y_i|X_i,U_i;\varepsilon_y)P(W_i|X_i;\varepsilon_\eta)P(X_i|U_i;\varepsilon_\alpha)\}. \quad (5)$$

We set these parameters β, ψ, $\alpha_k = (\alpha_{k0},\alpha_{kv}^T)^T$ for $k = 1,\dots,r$ and σ_z^2 with a priori obey the following distribution

$$\beta|\psi,\beta^0,H_\beta^0\sim N_{r+p}(\beta^0,\psi^{-1}H_\beta^0),\ \psi^{-1}|b_1,b_2\sim\Gamma(b_1,b_2),$$
$$\alpha_k|\alpha_k^0,\ H_{\alpha k}^0\sim N_{p+1}(\alpha_k^0,\psi^{-1}H_{\alpha k}^0),\ \sigma_z^{-2}|d_1,d_2\sim\Gamma(d_1,d_2).$$

These b_1,b_2, β^0,H_β^0, $\alpha_k^0,H_{\alpha k}^0$, $d_1$and $d_2$are hyperparameters, and assume that their values are given by a priori information. According to the joint probability density function given above and their priors, we can use the Bayesian method to make statistical inferences on the parameters $\varepsilon = \{\varepsilon_y,\varepsilon_\alpha,\varepsilon_\eta\}$. In addition, we use Gibbs sampling and Metropolis-Hastings algorithm to analyze the measurement error model with longitudinal data. We get the posterior distribution of the interested parameters

$$P(\sigma_z^{-2}|X,U,\alpha)\sim\Gamma(d_1+0.5n,d_2+0.5\sum_{k=1}^r\sum_{i=1}^n(X_{ki}-\alpha_{k0}-\alpha_{kv}^TU_i)^2);$$
$$P(\beta|\psi,Y,X,U)\propto exp\{\psi^{-1}\sum_{i=1}^n(Y_i\tau_i-d(\tau_i)),0.5\psi^{-1}(\beta-\beta^0)^T(H_\beta^0)^{-1}(\beta-\beta^0)\}P(\alpha_k|X,\ U,\sigma_x^2)\sim N(\mu_{\alpha_k}^*,\Omega_{\alpha_k}^*)$$

where

$$\mu_{\alpha_k}^{*} = \Omega_{\alpha_k}^{*}(\sum_{i=1}^{n} U_i^{*} X_i \sigma_z^{-2} + (H_{\alpha k}^{0})^{-1}\alpha_k^{0})\Omega_{\alpha_k}^{*} = (\sum_{i=1}^{n} U_i^{*} U_i^{*T}\sigma_z^{-2} + (H_{\alpha k}^{0})^{-1})^{-1},\ U_i^{*} = (1, U_i^{T})^{T}.$$

3. Simulation and Bayesian Estimation

In order to test the feasibility of the Bayesian method in the case where our previously assumed model obeys a large number of different distributions in the measurement error η_{ij}, for the sample size $n = 200, m = 5$ in the generalized linear measurement error model, 100 sets of data sets $\{(Y_i, U_i, W_i, X_i): i = 1, \ldots, n\}$ are repeatedly generated from the probability density function with (Bernoulli) distribution for simulation studies

$$Y_i \sim B(1, p_i). \tag{6}$$

Here $\lambda_i = log\{p_i/(1-p_i)\} = X_i^T\beta_z + U_i^T\beta_v = Z_i^T\beta$. Assumptions $U_i \sim N(0, 0.25I_3)$, X_{1i} and X_{2i} are derived from the data generated according to equation (4). Under this distribution, the value ψ is known to be a constant according to equation (1). For $k = 1$ and 2, the true values $\beta_z, \beta_v, \alpha_k$ and σ_z^2 are taken as

$\beta_z = (0.8, 0.9)^T, \beta_v = (0.5, 0.5, 0.5)^T$, $\alpha_k = (0.2, 0.2, 0.2, 0.5)^T$ and σ_z^2=1. To test the validity of the prior metric measurement error $\eta_{ij} = (\eta_{ij1}, \eta_{ij2})^T$ with TCDP (Truncate and Centered Dirichlet process), we consider the following two distribution hypotheses for η_{ijk}:

Simulation 1: We assume η_{ijk} that it is from $\eta_{ijk} \sim N(0, 1.2^2)$.

Simulation 2: We assume η_{ijk} that it is from

$$\eta_{ijk} \sim 0.6N(-0.4, 0.2^2) + 0.4N(0.6, 0.2^2).$$

Simulation 3: We assume η_{ijk} that it is from

$$\eta_{ijk} \sim 0.3N(0.5, 0.1) + 0.2N(3, 0.1) + 0.5N(-1.5, 0.1)$$

Simulation 4: We assume η_{ijk} that it is from

$$\eta_{ijk} \sim 0.3N(0.5, 0.1) + 0.2N(3, 0.1) + 0.1N(-3.5, 0.1) + 0.4N(-1, 0.1).$$

To study the sensitivity of a priori to Bayesian estimation, we consider the following three priori assumptions for the parameters β and α_k.

Type A. About a priori hyperparameters β and α_k are chosen to be β, β, β and β. This ensures a good priori information in the simulation test.

Type B. About a priori hyperparameters β and α_k are chosen to be $\beta^0 = 1.5 \times (0.8, 0.9, 0.5, 0.5, 0.5)^T$, $H_\beta^0 = 0.75I_5$, $H_{\alpha k}^0 = 0.75I_4$ and $\alpha_k^0 = 1.5 \times (0.2, 0.2, 0.2, 0.5)^T$. This ensures a weak priori information in the simulation test.

Type C. About a priori hyperparameters β and α_k are chosen to be $\beta^0 = 0 \times (0.8, 0.9, 0.5, 0.5, 0.5)^T$, $H_\beta^0 = 10I_5$, $\alpha_k^0 = 0 \times (0.2, 0.2, 0.2, 0.5)^T$ and $H_{\alpha k}^0 = 10I_4$. This ensures noninformative prior information in the simulation test.

The simulation results are listed in Table 1-4. We drop the first 5000 iterations of all parameters and collect 5000 data after 5000th to generate 100 sets of data from the posterior distribution of the full data through Marko. Bayesian Monte Carol (MCMC) sampling was used to evaluate Bayesian estimates. The results of the above two hypotheses and their three different a priori designs are given in Tables 1-4, where 'Bias' is the absolute value of the difference between the true value and the parameter mean of the 100 sets of replicates;

and 'RMS' It is the mean square error of the parameter estimates and true values for 100 replicates. We also have plotted densities of η_{ijk} and $\hat{\eta}_{ijk}$ for Simulation 4 under Type C prior inputs in Fig. 1.

Table 1. First Simulated Parameter Estimation

Parameter	True value	Type A Bias	A RMS	Type B Bias	B RMS	Type C Bias	C RMS
α_{10}	0.2	0.0032	0.0731	0.0164	0.0272	0.0059	0.0594
α_{11}	0.2	0.0439	0.1536	0.0047	0.1257	0.0038	0.1846
α_{12}	0.2	0.0076	0.1594	0.0367	0.1641	0.0531	0.1763
α_{13}	0.5	0.0061	0.1079	0.0282	0.1174	0.0066	0.1237
α_{20}	0.2	0.0138	0.0669	0.0180	0.0635	0.0337	0.0669
α_{21}	0.2	0.0128	0.1711	0.0135	0.1062	0.0717	0.1151
α_{22}	0.2	0.0078	0.1316	0.0287	0.1150	0.0418	0.1599
α_{23}	0.5	0.0269	0.1443	0.0481	0.1688	0.0191	0.1652
β_0	0.8	0.0173	0.0858	0.0332	0.0896	0.0202	0.0256
β_1	0.9	0.0034	0.0763	0.0060	0.0702	0.0310	0.0712
β_2	0.5	0.0374	0.1391	0.0854	0.1083	0.0236	0.1249
β_3	0.5	0.0356	0.1330	0.0418	0.1834	0.0265	0.1395
β_4	0.5	0.0132	0.1458	0.0229	0.1964	0.0335	0.1735
σ_z^2	1.0	0.0073	0.0805	0.0053	0.0652	0.0086	0.0347

Table 2. Second Simulated Parameter Estimation

Parameter	True value	Type A Bias	A RMS	Type B Bias	B RMS	Type C Bias	C RMS
α_{10}	0.2	0.0186	0.0352	0.0146	0.0756	0.0108	0.0863
α_{11}	0.2	0.0150	0.1308	0.0082	0.1488	0.0126	0.1395
α_{12}	0.2	0.0185	0.1991	0.0007	0.1526	0.091	0.1758
α_{13}	0.5	0.0008	0.1293	0.0315	0.1697	0.0359	0.1382
α_{20}	0.2	0.0064	0.0919	0.0326	0.0814	0.0227	0.0957
α_{21}	0.2	0.0003	0.1638	0.0415	0.1364	0.0244	0.1705
α_{22}	0.2	0.0388	0.1313	0.0226	0.1195	0.0008	0.1440
α_{23}	0.5	0.0089	0.1357	0.0475	0.1384	0.0397	0.1254
β_0	0.8	0.0046	0.0833	0.0245	0.0619	0.0047	0.0455
β_1	0.9	0.0062	0.0546	0.0293	0.0550	0.0132	0.0644
β_2	0.5	0.0137	0.1252	0.0426	0.1856	0.0325	0.2420
β_3	0.5	0.0601	0.1327	0.0446	0.3213	0.0618	0.1754
β_4	0.5	0.0160	0.1882	0.0201	0.1721	0.0080	0.1394
σ_z^2	1.0	0.0061	0.0624	0.0075	0.0689	0.0092	0.0686

Table 3. Third Simulated Parameter Estimation

Parameter	True value	Type A Bias	A RMS	Type B Bias	B RMS	Type C Bias	C RMS
α_{10}	0.2	0.0002	0.0482	0.0007	0.0425	0.0001	0.0436
α_{11}	0.2	0.0112	0.1152	0.0171	0.1489	0.0146	0.1323
α_{12}	0.2	0.0117	0.1221	0.0016	0.1144	0.0208	0.1539
α_{13}	0.5	0.0018	0.1372	0.0133	0.1494	0.0324	0.1442
α_{20}	0.2	0.0035	0.0856	0.0341	0.0817	0.0229	0.0886
α_{21}	0.2	0.0045	0.1108	0.0203	0.1554	0.0132	0.1662
α_{22}	0.2	0.0243	0.1030	0.0115	0.1059	0.0016	0.1400
α_{23}	0.5	0.0035	0.1138	0.0191	0.1233	0.0146	0.1235
β_0	0.8	0.0046	0.0324	0.0112	0.0309	0.0147	0.0626
β_1	0.9	0.0035	0.0603	0.0130	0.0796	0.0064	0.0982
β_2	0.5	0.0125	0.1340	0.0252	0.1697	0.0103	0.1236
β_3	0.5	0.0105	0.1243	0.0358	0.1452	0.0209	0.1637
β_4	0.5	0.0206	0.1338	0.0301	0.1277	0.0031	0.1458
σ_z^2	1.0	0.0135	0.0562	0.0243	0.0775	0.0095	0.0864

Table 4. Fourth Simulated Parameter Estimation

Parameter	True value	Type A Bias	RMS	Type B Bias	RMS	Type C Bias	RMS
α_{10}	0.2	0.0162	0.0562	0.0142	0.0766	0.0191	0.0864
α_{11}	0.2	0.0060	0.1386	0.0092	0.1685	0.0125	0.1574
α_{12}	0.2	0.0292	0.1133	0.0024	0.1142	0.0209	0.1476
α_{13}	0.5	0.0018	0.1814	0.0241	0.1792	0.0361	0.1549
α_{20}	0.2	0.0086	0.0411	0.0103	0.0614	0.0224	0.0881
α_{21}	0.2	0.0013	0.1192	0.0405	0.1543	0.0264	0.1753
α_{22}	0.2	0.0195	0.1320	0.0321	0.1286	0.0027	0.1405
α_{23}	0.5	0.0062	0.1116	0.0143	0.1236	0.0216	0.1330
β_0	0.8	0.0144	0.0162	0.0112	0.0619	0.0256	0.0662
β_1	0.9	0.0156	0.0703	0.0163	0.0551	0.0134	0.0844
β_2	0.5	0.0241	0.1526	0.0442	0.1895	0.0327	0.1124
β_3	0.5	0.0160	0.1249	0.0254	0.1399	0.0216	0.1768
β_4	0.5	0.0204	0.1135	0.0406	0.1183	0.0102	0.1821
σ_z^2	1.0	0.0079	0.0853	0.0041	0.0897	0.0052	0.0752

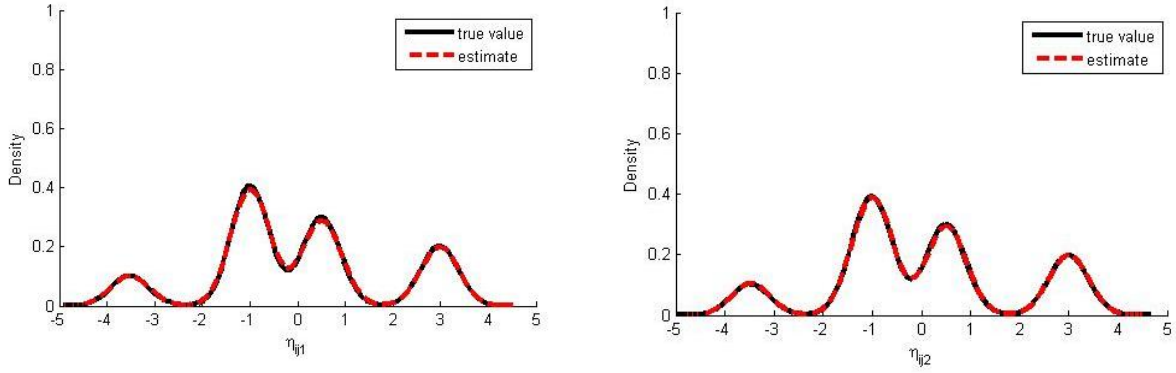

Fig. 1. Estimated versus true densities of $\eta_{ij1}$and η_{ij2} as assumption under Type C prior inputs.

4. Conclusion

According to Table 1-4 and Fig. 1, we know that (i) the model uses Bayesian estimation to be reasonable and correct, regardless of the distribution η_{ijk} and a priori assumptions, because the unknown parameters produce Bias values less than 0.1 and RMS values less than 0.2. (ii) Dirichlet priori is generally sufficient to capture the characteristics of the various distribution hypotheses of the measurement error model. (iii) The results show that the proposed method is a good estimate of the distribution η_{ijk}.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Meilan Qiu and Dewang Li conducted the research; Meilan Qiu and Zhongyi Ke analyzed the data; Dewang Li and Zhongyi Ke wrote the paper; all authors had approved the final version.

Acknowledgment

The project is supported by natural science foundation of Guangdong province of China (2018A030310038) and the talent project of Huizhou University of China (2017JB010), and by Social Science Fund of Guangdong province of China (GD16XYJ13), National Statistical Science Research Project of China (2017LY66).

References

[1] You, J. H., & Chen, G. M. (2006). Estimation of a semiparametric varying-coefficient partially linear errors-in-variables model. *Journal of Multivariate Analysis, 97(2),* 324-341.

[2] Zhou, Y., & Liang, H. (2011). Statistical inference for semi-parametric varying-coefficient partially models with error-prone linear covariates. *The Annals of Statistics, 37(1),* 427-458.

[3] Wei, C. H. (2010). Estimation of Error model with partially linear variable coefficient variables with missing variable. *Acta Mathematica Scientia Sinica, 30A(4),*1042-1054.

[4] Wei, C. H. (2012). Statistical inference for restricted partially linear varying-coefficient errors-in-variables models. *Journal of Statistical planning and Inference, 142(8),* 2464-2472.

[5] Liang, H., Hardel, W., & Carroll, R. J. (1999). Estimation in a semiparametric partially linear error-in-variables models. *The Annals of Statistics, 27*, 1519-1535.

[6] Ma, Y. Y., & Carroll, R. J. (2006). Locally efficient estimators for semiparametric models with measurement error. *J. Am. Statist. Assoc, 101,* 1465-1474.

[7] Liang, H., Wang, S. J., & Carroll, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika, 94,* 185-198.

[8] Pan, W. Q., Zeng, D. L., & Lin, X. H. (2008). Estimation in semiparametric transition measurement error models for longitudinal data. *Biometrica, 65,* 728-736.

[9] Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721-741.

[10] Lachos, V. H., Cancho, V. G., & Aoki, R. (2010). Bayesian analysis of skew-t multivariate null intercept measurement error model. *Stat Pap., 51,* 531-545.

Dewang Li was born in Anyuan, Jiangxi Province, China in January, 1976. He studied at Yunnan University. He obtained a probability theory Ph.D. in mathematics and mathematical statistics in 2015. He is mainly engaged in the fields of mathematical statistics and Bayesian statistics. He is a lecturer. He worked in the Department of Mathematics, Hechi College, Guangxi, China from 2008 to 2012. From 2015 to now, he works in the Department of Statistics, School of Mathematics and Statistics, Huizhou University, Guangdong, China. He has published SCI articles in communications in statistics - theory and methods, stat papers, and advances in mathematical physics.

Meilan Qiu was born in 1981, Jiangxi Province, Ganzhou city. Ph.D., a lecturer of Huizhou University. She mainly studies the theory of partial differential equations and complex flow coupling model as well as its numerical solution, profound understand the basic theory and programming calculation of the finite difference method, the finite element method (FEM) and LDG (local discontinuous finite element) method and so on. She can write the corresponding Matlab code to enhance the implementation capability and have a deep understanding of the design of numerical calculation method for complex coupled flow equations and its numerical theoretical analysis. At present, she is mainly engaged in modeling, theoretical analysis and numerical simulation of complex porous media flow and fluid coupling in large cracks or pipelines. She published 9 SCI papers, presided over one Natural Science Foundation Project of Guangdong

Province(2018A030310038) and participated in four projects of the National Natural Science Foundation of China.

Zhongyi Ke is a professor and master tutor. He was born in Huanggang City, Hubei Province, China in 1969. His research areas are mainly applied econometrics, innovation economics, and technological innovation games. He chaired the Social Science Fund of Guangdong Province of China (GD16XYJ13) and National Statistical Science Research Project of China (2017LY66).