Semiparametric Bayesian Estimation of Bernoulli Model with Measurement Error

Dewang Li[1], Meilan Qiu[1*], Yuanying Zhao[2*]

[1] School of Mathematics and Big Data, Huizhou University, Huizhou, Guangdong, 516007, China.
[2] College of Mathematics and Information science, Guiyang University, Guizhou, 550005, China.

* Corresponding author. Tel.: 86-18319629062, 86-18585034735; email: qml_1981@126.com, zhaoyuanying_@126.com
Manuscript submitted November 4, 2018; accepted May 1, 2019.
doi: 10.17706/ijapm.2019.9.4.167-172

Abstract: In this paper, we relax the fully parametric distributional assumption of measurement errors (MEs) to establish mixture Bernoulli model by a centered Dirichlet process. A hybrid algorithm is presented to generate observations required for a Bayesian inference from the posterior distributions of parameters and covariates subject to MEs in Bernoulli model by combining the stick-breaking prior and the Gibbs sampler together with the Metropolis-Hastings algorithm. Two Monte Carlo studies illustrate the superiority of the measurement error estimators in certain situations.

Key words: Monte Carlo, Bayesian estimation, Bernoulli model, measurement error.

1. Introduction

In a logistic regression model when covariates are subject to measurement error the naïve estimator, obtained by regressing on the observed covariates, is asymptotically biased. A measurement error model is a linear or non-linear regression model with measurement error in the explanatory variables. Disregarding these measurement errors in estimating the regression parameters results in asymptotically biased, i.e. inconsistent estimators. This is the motivation for investigating measurement error models. On the other hand, most studies cannot be recorded exactly in the life sciences, biology, ecology and economics involve variables. Recently measurement error methods have been applied in the masking of data to assure anonymity [1]. In engineering, the cabibration of measuring instruments deals with measurement errors by definition [2], many more examples and contribution to this field can be found in the literature, in particular in [3]-[5].

Due to the importance of the measurement error problems, there are huge amount of papers and several books on measurement errors. It is important for us to review relatively recent developments in econometrics and statistics literature on measurement error problems. Reviews of earlier results on this subject can be found in Fuller [3], Carroll, Ruppere and Stefanski [6], Wansbeek and Meijer [7], Bound, Brown and Matwiowetz [8], Hausman [9].

In this survey we aim at developing a semiparametric Bayesian approach to simultaneously obtain Bayesian estimations of parameters and covariates subject to MEs by combing the stick-breaking prior and Gibbs sampler together with the Metropolis-Hastings algorithm.

2. Generalized Linear Measurement Error Models

Suppose y_i denote the observed outcome variable, X_i be a $p \times 1$ vector of the unobserved covariate variables, and V_i be a $r \times 1$ vector of the observed covariate variables for the i th individual with $i = 1, \dots, n$. Giving $Z_i = (X_i^T, V_i^T)^T$, we consider that y_i's are conditionally independent of each other, the conditional probability density function of y_i is assumed by

$$p(y_i | Z_i, \lambda) = \exp\{\frac{y_i \gamma_i - d(\gamma_i)}{\lambda} + c(y_i, \lambda)\}. \tag{1}$$

with $\phi_i = E(y_i | Z_i) = d'(\gamma_i) = \frac{\partial d(\gamma_i)}{\partial \gamma_i}$ and $U_i = Var(y_i | Z_i) = \lambda d''(\gamma_i) = \frac{\partial^2 d(\gamma_i)}{\partial \gamma_i^2}$, where λ is a scale parameter, $d(\cdot)$ and $c(\cdot,\cdot)$ are specific differentiable functions. The conditional mean ϕ_i is given to satisfy

$$\eta_i = h(\phi_i) = X_i \rho_x + V_i^T \rho_u = Z_i^T \rho. \tag{2}$$

where $h(\cdot)$ is a monotonic differentiable link function, $\rho = (\rho_x, \rho_u^T)^T$ is a $(p + r) \times 1$ vector of unknown regression coefficients. If the true covariate X_i are measured m times for individual i, giving outcomes W_{ij} for $j = 1, \dots, m$ the structural ME model can be defined as

$$W_{ij} = X_i + \theta_{ij} \quad . \tag{3}$$

where the MEs θ_{ij}'s are assumed to follow an unknown distribution, and are independent of X_i. Following Lee *et al.* [10], we assume the Dirichlet process (DP) mixture model to specify the distribution of θ_{ij}.

The true covariate model for X_i can be defined as

$$X_i = \beta_0 + \beta_u^T V_i + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma_x^2). \tag{4}$$

where β_0 is an intercept, $\beta_u = (\beta_1, \dots, \beta_r)^T$ and $\rho_u = (\rho_1, \dots, \rho_r)^T$ is a $r \times 1$ vector of unknown regression parameters. Let $Y = \{y_1, \dots, y_n\}$, $X = \{X_1, \dots, X_n\}$, $V = \{V_1, \dots, V_n\}$, $\theta = \{\theta_1, \dots, \theta_n\}$ and $W = \{W_1, \dots, W_n\}$ in which, $\theta_i = (\theta_{i1}, \dots, \theta_{im})$ and $W_i = \{W_{i1}, \dots, W_{im}\}$ for $i = 1, \dots, n$. We let $\alpha_y = \{\rho, \lambda\}$, α_θ are parameters of equation (3) and $\alpha = \{\alpha_y, \alpha_\beta, \alpha_\theta\}$. The joint probability density function for $\{Y, W, \theta, X\}$ is given by

$$P(Y, W, \theta, X | V, \alpha) = \prod_{i=1}^{n} \{p(y_i | X_i, V_i; \alpha_y) p(W_i | X_i; \alpha_\theta) p(X_i | V_i; \alpha_\beta)\}. \tag{5}$$

a_1, a_2, ρ^0, H_ρ^0, β^0, H_β^0, c_1 and c_2 are hyperparameters whose values are considered to be given by the prior information. We assume the following priors for parameters ρ, λ, $\beta^* = (\beta_0, \beta_u^T)^T$ and σ_x^2:

$$\rho \Big| \lambda, \rho^0, H_\rho^0 \sim N_{r+p}(\rho^0, \lambda^{-1} H_\lambda^0), \quad \lambda^{-1} \Big| a_1, a_2 \sim \Gamma(a_1, a_2),$$

$$\beta^* \Big| \beta^0, H_\beta^0 \sim N_{v+1}(\beta^0, \lambda^{-1} H_\beta^0), \quad \sigma_x^{-2} \Big| c_1, c_2 \sim \Gamma(c_1, c_2).$$

Using the above presented joint probability density function and priors, we develop the generalized linear measurement error models. At the same time, utilizing the Gibbs sampler together with the Metropolis-Hastings algorithm for our defined models, we make statistical inference on parameters in $\theta = \{\theta_y, \theta_\rho, \theta_\xi\}$ with a Bayesian approach.

3. Bernoulli Simulation and Bayesian Estimations

We consider data that are composed of a response and a covariate X_i for $i = 1, \ldots, n$. We define the Bernoulli distribution $B(1, p_i)$ with $\eta_i = \log \frac{p_i}{1 - p_i} = X_i \rho_x + V_i^T \rho_u = Z_i^T \rho$. Let $V_i \sim N(0, 0.25 I_3)$ and X_i is generated via Equation (4). In this case, λ relating to Equation (1) is a constant. The true values of ρ_x, ρ_u, β and σ_x^2 are taken to be ρ_x=0.9, $\rho_u = (0.6, 0.6, 0.6)^T$, $\beta = (0.3, 0.3, 0.3, 0.6)^T$ and σ_x^2=1 for $n = 100$, $m = 4$. To investigate the effectiveness of our proposed methods, we consider the following distributional assumption for θ_{ij}

Assumption 1: We assume the distribution of $\theta_{ij} \sim N(0, 1.1^2)$.

Assumption 2: We assume the distribution of $\theta_{ij} \sim 0.6 N(-0.4, 0.2^2) + 0.4 N(0.6, 0.2^2)$.

In order to inspect sensitivity of Bayesian estimates by different prior inputs, we select the following three types of priors for ρ and β.

Type A. The hyperparameters corresponding to the priors of ρ and β are chosen to be $\rho^0 = (0.9, 0.6, 0.6, 0.6)^T$, $H_\rho^0 = 0.25 I_4$, $\beta^0 = (0.3, 0.3, 0.3, 0.6)^T$ and $H_\beta^0 = 0.25 I_4$. This can be regarded as a situation with good prior information.

Type B. The hyperparameters corresponding to the priors of β and ρ_k are taken to be $\rho^0 = 1.5 \times (0.9, 0.6, 0.6, 0.6)^T$, $H_\rho^0 = 0.75 I_4$, $\beta^0 = 1.5 \times (0.3, 0.3, 0.3, 0.6)^T$ and $H_\beta^0 = 0.75 I_4$. This can be regarded as a situation with inaccurate prior information.

Type C. The hyperparameters corresponding to the priors of β and ρ_k are taken to be $\rho^0 = 0 \times (0.9, 0.6, 0.6, 0.6)^T$, $H_\rho^0 = 10 I_4$, $\beta^0 = 0 \times (0.3, 0.3, 0.3, 0.6)^T$ and $H_\beta^0 = 10 I_4$. This can be regarded as a situation with noninformative prior information.

After 10000 burn-in iterations 5000 observations are collected in each of the generated 100 data sets, we evaluate Bayesian estimates via Markov chain Monte Carol (MCMC) samples from the full data posterior distribution. Results of Table 1-2 are presented under assumption together with three types of prior inputs. In Table 1-2, `Bias' is the absolute difference between the true value and the mean of the estimates based on 100 replications and `RMS' is the root mean square between the estimates based on100 replications and its true value.

Table 1. Parameter Estimates in the First Simulation

Parameter	True value	Type A Bias	Type A RMS	Type B Bias	Type B RMS	Type C Bias	Type C RMS
β_0	0.3	0.0103	0.0742	0.0031	0.0789	0.0088	0.0744
β_1	0.3	0.0172	0.1918	0.0055	0.1872	0.0026	0.2181
β_2	0.3	0.0105	0.2045	0.0421	0.1777	0.0083	0.2160
β_3	0.6	0.2988	0.3723	0.0022	0.2093	0.0008	0.1760
ρ_x	0.9	0.0164	0.2744	0.1194	0.3725	0.0738	0.5163
ρ_1	0.6	0.0189	0.2610	0.0630	0.3616	0.0108	0.4539
ρ_2	0.6	0.0332	0.2584	0.1334	0.3855	0.0283	0.5220
ρ_3	0.6	0.0009	0.1658	0.0512	0.1774	0.0582	0.1907
σ_z^2	1.0	0.1058	0.1555	0.0657	0.1387	0.0897	0.1520

Table 2. Parameter Estimates in the Second Simulation

Parameter	True value	Type A Bias	Type A RMS	Type B Bias	Type B RMS	Type C Bias	Type C RMS
β_0	0.3	0.0065	0.0738	0.0075	0.0698	0.0025	0.0637
β_1	0.3	0.0042	0.1888	0.0085	0.2028	0.0145	0.1979
β_2	0.3	0.0278	0.2019	0.0224	0.1994	0.0001	0.1903
β_3	0.6	0.0336	0.1900	0.0050	0.2052	0.0251	0.1705
ρ_x	0.9	0.0267	0.2706	0.1094	0.3969	0.0260	0.5386
ρ_1	0.6	0.0011	0.2144	0.0669	0.4128	0.0146	0.4630
ρ_2	0.6	0.0119	0.2398	0.0636	0.3576	0.0339	0.4945
ρ_3	0.6	0.0146	0.1452	0.0110	0.1524	0.0002	0.1559
σ_z^2	1.0	0.0270	0.0924	0.0441	0.1046	0.0578	0.1050

4. Conclusion

Results from Tables 1-2 shows that 1) even if the different distributional assumptions of θ_{ij} and prior inputs of unknown parameters, Bayesian estimates the Bernoulli model with measurement error are reasonably accurate because their Bias values were less than 0.10 and their RMS values were less than 0.20; 2) using our proposed method, we can estimate the mean and standard deviation of the true distribution of θ_{ij} well; 3) the performance of the proposed procedures is developed in the Bernoulli model with measurement error.

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, 2015JB018). The research is also supported by grants from the National Natural Science Foundation of China (11761016), and by the special funding of Guiyang science and technology bureau and Guiyang University (GYU-KYZ[2018]04).

References

[1] Brand, R. (2002). Microdata protection through noise addition in Ference control in statistical databases-FROM Theory to practice. *Lecture Notes in Computer Science 2316*. Berlin: Springer.

[2] Brown, P. J. (1982). Multivariate calibration. *Journal of the Royal Statistical Society-Series B, 44*, 287-321.

[3] Fuller, W. (1987). *Measurement Error Models*. New York: Wiley.

[4] Cheng, C. L., & Ness, L. W. (1999). *Statistical Regression with Measurement Error*. London: Around.

[5] Carroll, R. J., Ruppert, D. Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. London: Chapman & Hall.

[6] Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models: A Modern Perspective.* London: Chapman & Hall.

[7] Wansbeek, T., & Meijer, E. (2000). *Measurement Error and Latent Variables in Econmetrics*. New York: North Holland.

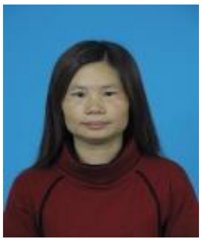
[8] Bound, J., Brown, C., & Mathiowetz, N. (2001). *Measurement Error in Survey Data, in Handbook of Econometrics.* New York: North Holland.

[9] Hausman, J. (2001). Mismeasured variables in econometric analysis: problems from the light and problems from the left. *The Journal of Economic Persoectives, 15*, 57-67.

[10] Lee, S. Y., Lu, B., & Song X. Y. (2008). Semiparametric Bayesian analysis of structural equation models with fixed covariates. *Statistics in Medicine, 15*, 2341-2360.

Dewang Li was born in Anyuan, Jiangxi province, China in January 1976. He studied at Gannan Normal University, Yunnan University and Yunnan University. He obtained a bachelor of science in mathematics education (1999); a master of science in probability and mathematical statistics (2008) and a probability theory; Ph.D. in mathematics and mathematical statistics (2015). He is mainly engaged in the fields of mathematical statistics and Bayesian statistics.

He is a lecturer, 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 Big Data, 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. 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, he 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.

Zhao Yuanying was born in 1981. He is an associate professor of Guiyang College. He is the seventh director of the Guizhou Provincial Statistical Research Association. The main research areas are complex data analysis and statistical learning methods. He has presided over one National Natural Science Foundation project, one Guizhou Science and Technology Fund, and participated in two projects of the National Natural Science Foundation of China, one National Social Science Fund Project, two national statistical science research projects and one Natural Science Foundation of the Education Department.