Innovations

Bayesian Estimation of Correlated Error Variables in Binary Logistic Model

Onatunji Adewale. Paul[1], Binuyo Adeyemi Olukayode[2], Olalude Oladapo A[3]., Amzat Kafayat Temitope[4] and Ajao Olutunde Michael[5]

[1] Department of Statistics, Faculty of Pure and Applied Sciences, Ladoke AkintolaUniversity Technology, Ogbomoso, Oyo state, Nigeria.

[2,4]Department of Mathematical Sciences, Faculty of Natural Sciences, Ajayi Crowther Universisty, Oyo, Oyo State, Nigeria.

[4]Departement of Mathematical Sciences, College of Natural and Applied Sciences, Crescent University, Abeokuta, Ogun State, Nigeria.

Corresponding author: **Onatunji Adewale. Paul**

Abstract

Logistic model used for epidemiological studies usually covers binary response with covariates contaminated with error and collinearity. The regression violations (measurement error and collinearity) in no small affect statistical model making estimates statistically fallacious(bias, wide confidence interval, incorrect interpretation of hypothesis testing, convergence problem). This study is aimed at proposing Bayesian approach to jointly investigate the effect of varying degree of contaminated normal distribution in covariate under linear collinearity. The error variables with reliability values(0, 1) were linearly correlated. The results show that the Bayesian estimates are performing well at high varying degree of reliability values in support of the report of Bartlett and Keogh(2018). The posterior estimates are however not unbiased at fixed sample size where the covariates are incorrectly specified and correlated. Conclusively, the acceptably prefered reliability value for Bayesian estimation of logistic model with basic violations is 0.7.

Keywords: *reliability values, collinearity, measurement error, Bayesian approach, logistic model*

Introduction

Regression assumptions play a vital role in the results obtained from logistic method of regression. However, when the basic assumptions of logistic regression are violated(measurement error and multicollinearity), the consequence is grievous and that there is high likelihood of making statistically insignificant conclusion from the analysis. In the epidemiological studies, the use of Logistic regression does not only involve binary response but also correlated covariates caused by measurement errors. This regression modelwith which outcome of interest is inefficiently measured is then incorrectly specified due to presence of

multicollinearity. This multicollinearity is a statistical happening when a set of variables is ill conditional if there is a linear relationship among continuous independent variables. The effects of it in all regression methods with no exception to logistic regression does not only cause biased but also unreliable estimates with large standard errors.In addition, continuous independent variables in many study are often incorrectly measured with error and multicollinearity. Error in the covariate in the logistic regression causes biasedness. This error is the variation between observed and expected measurement, divided into systematic and random errors. Thus, multiple covariates with any of forms of the errors are unidentifiable (Kuchenhoff, 1995). It is not clear how measurement errors affect regression results(Junhan and Grace,2020).On the contrary, Clark(1982) reported that logistic regression of estimator of β_0 is not asymptotically unbiased. This causes high likelihood of incorrect hypothetically interpretation of results.

However, Bayesian in recent times has gained a wide attention in dealing with regression violations. Hyang-Mi and Burstyn (2009) showed that Bayesian method improves estimates of odds ratios when errors are additively measured in the outcomes. The use of conditional logistic regression likelihood as a disease model for several mismeasured continuous exposures under a classical measurement error model was successfully handled by proposed Bayesian method (Espino-Hernandez et .al, 2011). Mwalili et.al (2005) applied Bayesian method to correct for interobserver measurement error in an ordinal logistic regression model. The Bayesian method is flexible in handling both covariate measurement error and missing data to produce unbiased estimates when the reliability of error prone covariates is high, but otherwise when it is low(Bartlett and Keogh,2018). In this paper, Markov Chain Monte Carlo(MCMC) used facilitated the propagation of measurement errors to estimate the logistic model parameters considerably.

The work aims at jointly investigating the combined effects of measurement error and multicollinearity in logistic regression using Bayesian approach and, is organized as follows. Section 1 contains introduction, materials and method are contained in section two, section three discussion of results and finally section four contains conclusion of the paper.

Method 1: presentation of covariate measurement error model designs(CMEMD)

Few CMEMD developed in literature can be seen in Stefanski and Carroll(1985), Lee and Wilhelm(working paper CWP51/18), Kuchenhoff(1995). The framework of covariate measurement error model adopted is as follows

$$X_{MEV} = x + v, \text{ where } v = \sigma \sum^{1/2} \varepsilon_i \qquad (1)$$

Where $v = \sigma \sum^{1/2} \varepsilon_i$ and $\sum^{1/2}$ is the square root of a symmetric positive semidefinite matrix $\sum$ follows multivariate distribution, scaled so that $|\sum| = 1$ and ε_i which is independent of y_i, the response variable, is independently and identically random vector with mean zero and identity covariance. The scaled factor $\sigma \sim N(0, d)$ with $d(0,1)$, is a contaminated normal distribution allowed in X_{MEV}, a vector of measurement error variable(MEV)in equation(1)

Method 2: presentation of correlated error variable model designs(CEMD)

Correlation between error metric covariates in epidemiological studies is common. It is, however, violates regression assumption. Its effects are no small way affect the results of regression and interpretation of results. There are many scenarios of collinearity/multicollinearity (see McDonald and Galarneau , 1975; Vatcheva et al., 2016; Yoo et al., 2014).When two MEVs are correlated it is called collinearity. If variance of X_{MEV} and γ change, then correlation coefficient(ρ) changes. Let $\gamma = \frac{std(v)}{std(X_{MEV})}$ for vector of covariate measurement errors. The correlation coefficient between two error variables, X_{1MEV} and X_{2MEV} is then given as

$$\rho = \frac{Cov(X_{1mev}, X_{2mev})}{\sqrt{Var(X_{1mev})}\sqrt{Var(X_{2mev})} + \sqrt{Var(v)}} \quad (2)$$

$$= \frac{\lambda}{(1+\gamma^2)^{1/2}} \quad (3)$$

Model

Binary logistic model is commonly used in epidemiological study when the response variable is dichotomous as alternative to ordinary least regression model(see Bartlett and Keog 2018; Fang and Yi , 2020;). The outcome model for correlated error explanatory variables is then defined as

$$\text{logit}\left(\text{P}\left(\text{Y}_\text{i} = 1/ X_{MEV1i}, X_{MEV2i}\right)\right) = \theta_0 + \theta_1 X_{MEV1i} + \theta_2 X_{MEV2i} \quad (4)$$

Bayesian estimation of correlated error variables of binary logistic regression

Bayesian estimation follows three patterns namely, likelihood function, prior and posterior distribution. In logistic regression the probability of success is different from one subjects to another depending on the vector of covariates(X_{iMEV} $i = 1, 2, ..., n$).

Then likelihood function of this subject follows binomial distribution given below

$$p(Y_i) = \pi\left(X_{iMEV}\right)^{Y_i}\left(1 - \pi\left(X_{iMEV}\right)\right)^{(1-Y_i)} \quad (5)$$

Where $Y_i = 1$ represent presence and $Y_i = 0$ represents absence of the event for the subject and $\pi(X_{MEVi})$ is the probability of event for that subject. The event of the subject follows a logistic distribution, $P(Y=1) = F(\theta_0 + \theta_1 X_{MEV1i} + \theta_2 X_{MEV2i})$ when F represents cdf of logistic distribution given below.

$$\pi(X_{MEVi}) = \frac{e^{\theta_0 + \theta_1 X_{MEV(1i)} + \theta_2 X_{MEV(2i)}}}{1 - e^{\theta_0 + \theta_1 X_{MEV(1i)} + \theta_2 X_{MEV(2i)}}} \quad (6)$$

The likelihood to i^{th} subject is thus given as

$$p(Y_i)=\left(\frac{e^{\theta_0+\theta_1 X_{MEV(1i)}+\theta_2 X_{MEV(2i)}}}{1-e^{\theta_0+\theta_1 X_{MEV(1i)}+\theta_2 X_{MEV(2i)}}}\right)^{Y_i}\left(\frac{e^{\theta_0+\theta_1 X_{cmev(1i)}+\theta_2 X_{cmev(2i)}}}{1-e^{\theta_0+\theta_1 X_{cmev(1i)}+\theta_2 X_{cmev(2i)}}}\right)^{1-Y_i} \quad (7)$$

Then, the individual subjects assume independently from each other over several subjects.

$$p(Y)=\prod_{i=1}^{n}\left(\frac{e^{\theta_0+\theta_1 X_{MEV(1i)}+\theta_2 X_{MEV(2i)}}}{1-e^{\theta_0+\theta_1 X_{MEV(1i)}+\theta_2 X_{MEV(2i)}}}\right)^{Y_i}\left(\frac{e^{\theta_0+\theta_1 X_{MEV(1i)}+\theta_2 X_{MEV(2i)}}}{1-e^{\theta_0+\theta_1 X_{MEV(1i)}+\theta_2 X_{MEV(2i)}}}\right)^{1-Y_i} \quad (8)$$

Prior specification for logistic regression parameters

Normal priors are used for this model parameters with zero mean and $\sigma^2=100$

$$p(\theta_j)\sim N(\mu_j,\sigma_j^2) \quad (9)$$

Posterior distribution

Posterior simulation was used to obtain conditional probability by multiplying likelihood function by prior specification in the study.

$$p(\theta/y_i)=\prod_{i=1}^{n}\left(\frac{e^{\theta_0+\theta_1 X_{MEV(1i)}+\theta_2 X_{MEV(2i)}}}{1-e^{\theta_0+\theta_1 X_{MEV(1i)}+\theta_2 X_{cmev(2i)}}}\right)^{y_i}\left(\frac{e^{\theta_0+\theta_1 X_{MEV(1i)}+\theta_2 X_{MEV(2i)}}}{1-e^{\theta_0+\theta_1 X_{MEV(1i)}+\theta_2 X_{MEV(2i)}}}\right)^{1-y_i}\times\prod_{j=0}^{p}\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{1}{2}\left(\frac{\theta_j-\mu_j}{\sigma_j}\right)^2\right) \quad (10)$$

Equation 10 does not have closed form expression. Therefore, posterior simulation is performed using Random walk Metropolis Hasting sampling for marginal posterior distribution for each coefficient a s implemented within STATA package.

Simulation Study

Marginal posterior distributions for the coefficients of correlated measurement error variables after setting true values, 0.2, 1, 1for θ_0, θ_1 and θ_2 respectively equation with fixed sample size(N) of 1000were obtained since posterior distribution is analytically impossible. Correlated error variables were generated using equation 1 with which a vector of covariates followed multivariate distribution contaminated with normal distribution of varying reliability values, $d(0,1)$ under scenario of linear collinearity, $X_{MEV1i}=X_{1MEV}$, $X_{MEV2i}=4X_{1MEV}-2X_{2MEV}$.MCMC sample size used is 20000 with Burn in of 5000.

The presence of measurement error in explanatory variables was tested using Delgado and Manteiga test. The study fails to reject no presence of measurement error for confidence level with p-value<0.05.

Convergence of MCMC sizes related to Effective Sample Size(ESS) and efficiency were obtained.

$$ESS=\frac{T}{\left(1+2\sum_{j=1}^{max\_lags}\rho_j\right)} \quad (11)$$

$$\rho_j=\frac{\gamma_j}{\gamma_0} \quad (12)$$

Here ρ_j is the lag-j of the autocorrelation of the MCMC sample size

Results

Table1.Posterior estimates of correlated error variables of Binomial logistic model with noninformative prior at fixed sample size(N=1000) and $d(0,1)$

$d(0,1)$	P-value	X_{MEV1}			X_{MEV2}		
		Pst.coef	Pst.std	MSCE	Pst.coef	Pst.std	MSCE
0.0	0.324	1.005216	8.932744	0.205395	1.512312	4.469503	0.102265
0.1	0.000	5.948465	0.4931269	0.012203	-1.077713	0.1489581	0.00359
0.2	0.000	5.306838	0.427706	0.012725	-0.9515292	0.1131923	0.003156
0.3	0.000	4.777907	0.3603794	0.009977	-0.854306	0.0916015	0.002339
0.4	0.000	4.354244	0.335229	0.009763	-0.7761093	0.0818004	0.00249
0.5	0.000	3.992	0.3032941	0.007482	-0.7100066	0.0735983	0.001936
0.6	0.000	3.691274	0.2806427	0.0076	-0.6571733	0.0661651	0.001748
0.7	0.000	3.440212	0.2521412	0.005533	-0.6126197	0.0593068	0.00137
0.8	0.000	3.216557	0.2395166	0.005864	-0.5725749	0.0554787	0.001297
0.9	0.000	3.032089	0.2224124	0.00513	-0.5395425	0.0513679	0.001193
1.0	0.000	2.866385	0.2189512	0.005911	-0.5094746	0.0500789	0.001426

Table1 shows the Bayesian estimates of Binomial logistic model when the variables are incorrectly measured $d(0,1)$ over collinearity. Existed in this study are two contaminated normally distributed random variables, X_{MEV1} and X_{MEV2} that are linearly correlated, with each having true parameter value of 1. Presence of measurement error in the explanatory variables was detected at p-value<0.05 for d (0.1,0.2,...,1.0).The obtained posterior estimates for X_{MEV1} ($X_{X_{MEV2}}$) are 5.948465(-1.077713), 3.992(-0.7100066), 3.032089(-0.5395425) and that of MSCE for X_{MEV1} ($X_{X_{MEV2}}$) are 0.012203(0.00359), 0.007482(0.001936), 0.00513(0.001193) for d , 0.1, 0.5 and 0.9 respectively. Posterior means of both X_{MEV1} and X_{MEV2} are not close to the true parameter value; but the MSCE considerably reduce at varying degree of incorrectness in variables under collinearity, except when d =1.However, the MSCE obtained when d =0.9 is small compared to others under collinearity

Table2. Convergence diagnostic of MCMC for model parameters after Bayesian estimation of correlated error variables of logistic model.

$d(0,1)$	P-value	X_{MEV1}		X_{MEV2}	
		ESS	EFFICIENCY	ESS	EFFICIENCY
0.0	0.324	1891.43	0.0946	1910.15	0.0955
0.1	0.000	1632.94	0.0816	1717.12	0.0859
0.2	0.000	1129.73	0.0565	1286.13	0.0643
0.3	0.000	1304.81	0.0652	1533.70	0.0767
0.4	0.000	1178.94	0.0589	1079.29	0.0540
0.5	0.000	1643.39	0.0822	1445.36	0.0723
0.6	0.000	1363.61	0.0682	1432.70	0.0716
0.7	0.000	2076.67	0.1038	1874.55	0.0937
0.8	0.000	1668.43	0.0834	1830.54	0.0915
0.9	0.000	1879.51	0.0940	1854.04	0.0927
1.0	0.000	1372.04	0.0686	1233.23	0.0617

Table 2 contains the effective sample size(ESS) within MCMC for Bayesian estimation of logistic regression model under basic violations: measurement error and linear collinearity. The obtained ESS for X_{MEV1} ($X_{X_{MEV2}}$) are 1632.94(1717.12), 2076.67(1874.55), 1372.04(1233.23)with that of efficiency are 0.0816(0.0859), 0.1038(0.0937) and 0.0686(0.0617) for reliability values(d), 0.1, 0.7 and 1.0 respectively. This implies that for all considered d (0,1) for linearly correlated variables, at least 1000 independent observation for each of the coefficients are within MCMC size. However, posterior estimates obtained when the independent variables are incorrectly measured at d =0.7 with efficiency >0.01 , X_{MEV1} (0.1038) and X_{MEV2} (0.0937) which has higher ESS as compared to others in this study are preferably acceptable.

Conclusion

The presence of measurement error and collinearity between covariates are major threats to regression modeling especially in epidemiological studies. Some of the effects of these violations are biased estimation, wrong hypothetical interpretations and unreasonable confidence intervals. The obtained results shows that at reliability value of 0.9, the MSCE is very low compared to others. Therefore, Bayesian method for estimating the effects of correlated error variable in logistic model at varying high degrees of reliability value 0.7, 0.8, 0.9 and 1.0 performs better. Also, the simulated error covariates using MCMC method were correlated at 0.7 for all considered reliability values.

References

1. *Kuchenhoff H.(1995). The identification of logistic regression model with errors in the variables.Statistical papers 36, 41-48*
2. *Clark R.R.(1982). The error-in-variables problem in the logistic regression model. Unpublished Ph.D thesis, university of North Carolina,Chapel Hill.*
3. *Hyang-Mi K. and Igor B.(2009) Bayesian Method for Improving Logistic Regression Estimates under Group-Based Exposure Assessment with Additive Measurement Errors, Archives of Environmental & Occupational Health, 64:4, 261-265,*
4. *Mwalili S. M., Lesaffre E, and Declerck D,(2005)A Bayesian ordinal logistic regression model to correct for inter observer measurement error in a geographical oral health study Appl. Statist..54, Part 1, pp. 77–93*
5. *Espino-Hernandez G., Gustafson P. and Burstyn I.(2011)Bayesian adjustment for measurement error in continuous exposures in an individually matched case-control study. MC Medical Research Methodology, 11:67*
6. *Stefanski L.A.and Carroll R.(1985).Covariate measurenment error in logistic regression.Theannal of Statistics. Vol.13, No.4 1335-1351*
7. *Vatcheva K. P., Lee M., McCormick J. B. and Rahbar M. H.(2016).Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiol,6:2*
8. *Lee and Wilhelm(working paper CWP51/18). Testing for the presence of measurement error in Stata, center for microdata and practice. The Institute for Fiscal Studies Department of Economics, UC.*
9. *Fang J.and Yi G.Y.(2020).Matrix-variate logistic regression with measurement error. Biometrika Trust. Page1-14.*
10. *Bartlett J. W and Keogh R. H.(2018).Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. Statistical Methods in Medical Research, Vol. 27(6) 1695–1708*
11. *Yoo W., MayberryR., Bae Sejong, H. Q. and Lillard J. W.(2015).A study of effects of multicollinearity in the multivariable analysis.Int J ApplSci Technol.*
12. *McDonald G. C. and Diane I. G.(1975). Monte Carlo Evaluation of Some Ridge. Type Estimator, Journal of the America Association, 70:350-416.*