International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Comparison of the Modified Two-Part Bayesian Linear Regression Model with Existing Approaches for Estimating Error in Small Samples

Samuel Joel Kamun

Catholic University of Eastern Africa
samuelkamun@gmail.com

ABSTRACT

The study compared the Modified Two-Part Bayesian Linear Regression Model in small samples for response-selective observations using other measurement error correction methods such as classical, systematic, Simex Homoscedastic, Simex Heteroscedastic, and Bayesian techniques.

Keywords: Bayesian Linear Regression, estimating error, small samples, standard error, sample bias

I Introduction

In biological research, measurement error poses a serious problem since it can lead to biased and inaccurate covariance estimations. It may be brought on by self-reporting, data coding errors, inaccurate or malfunctioning equipment, or the use of single measures in longitudinal processes. Measurement error evaluation becomes more pertinent as non-research data proliferates (Kamun, S. J., 2022). Measurement errors have not received adequate attention in applied epidemiology and medicine investigations, despite the abundance of resources devoted to covariate measurement errors. Only 28% of the 57 papers in the survey assessed its influence qualitatively, and only one of them measured it (Kamun, S. J., 2024).

II Review of Literature

The study examines the impact of childhood exposure to low electromagnetic fields on illness risk, considering measurement errors in regression models in epidemiologic research (Brazzale et. al. (2008)).

Error-prone variables include biological variation, assay volatility, and transmission error. All coefficients are subject to bias, even if the measurement error only affects individual interaction factors. Specific variables linked to both the result and the prone-to-error covariate can develop bias. The direction of bias is uncertain when a regression function's model has many covariates (Rosner et al., 1992; Carroll et. al., 1991).

The study uses regression calibration for response-selective observation on small samples of n = 13. It demonstrates that the classical measurement error correction algorithm has a lower variance when correcting for measurement error compared to other approaches, resulting in more efficient outcomes, despite the need for knowledge of the magnitude of measurement error (Rosner et al., 1992).

III Function for Generating Data

We suggest that by performing a series of operations on data according to a model:

$$f(y \mid x; \theta)g(x) \qquad (1)$$

we can produce or create data, where y is a response variable which is multivariate and x is a continuous or discrete vector of covariate variables and

$$f(y \mid x; \theta) \qquad (2)$$

is the regression part of the model. The marginal distribution of x is denoted by g(x) which for this study we have used Gaussian density to represent, is as shown below

$$K\left(u\right)=\frac{1}{\sqrt{\left(2\pi\right)}}e^{\left(\frac{-u^{2}}{2}\right)} \qquad (3)$$

Where $u = \left(\frac{x_1 - \bar{x}_1}{s} \right)$ and s is the standard deviation of x_1.

We estimate the conditional distribution of y for situations where there is no association with x_1. We describe this conditional distribution of y given x_1 as θ. When we take a small sample of n observations from the joint distribution of (y, x) or conditionally, when we sample all or some of the variables of x, then the necessary help to the main activity of the model, i.e., produce or create data, is given by x. We can also base our inference on the likelihood about θ.

The likelihood is given by

$\prod f(y \mid x; \theta)$ (4)

Since the probability of observation involves both (y, x), then there is need for the processes of estimation that is not dependent on the modeling of g(x) parametrically (Kamun, S. J. (2024)).

IV Selecting and Comparing Small Sample Sizes

The study looked and analyzed increasing sample sizes ranging from eight to twenty by examining their R squared values, bias, BIC, AIC, and standard error. By looking and carefully comparing only the R squared value, its corresponding bias, BIC, AIC and standard error, it appears that we could select the appropriate sample size for our study (Kamun, S. J. (2024)).

V The Modified Two-Part Model

The OLS model needs help to accurately model measurement error in a sample due to the difference between true exposure and replicated mismeasured exposure.

The modified two-part model considers replicated mismeasured exposure measures and their distribution-weighted properties, focusing on the probability of mismeasured exposure and fitting Bayesian distribution data conditioned on it.

For an exact solution suppose:

$Y = \beta_0 + \beta_X X + \beta_Z Z + \epsilon$ .... (5)

and

$X* = \alpha_0 + \alpha_X X + \alpha_Z Z + U$ .... (6)

Then

$E[Y \mid X*, Z] = E_{X|X*,Z}[E(Y \mid X*, Z) \mid X] = E_{X|X*,Z}[E(Y \mid Z, X)] = E_{X|X*,Z}[\beta_0 + \beta_X X + \beta_Z Z] = \beta_0 + \beta_X E[X \mid X*, Z] + \beta_Z Z$ ………….. (7)

We then regress Y on E[X | X∗, Z] and Z to get the right β coefficients. Then E[X | X∗, Z] is called the calibrated exposure.

Data is needed to estimate E[X | X∗, Z]. We use a validation subset where we observe the true X in an individual's subset.

Using measurement error and validation subset.

$X* = X + U$ .... (8)

Consider gamma approximation for distribution of (X, X∗):

$E[X \mid X*] = \mu_X + \frac{\text{cov}(X, X*)}{\text{var}(X \mid)} (X* - \mu_X) = \mu_X + \frac{\text{var}(X)}{\text{var}(X*)} (X* - \mu_X) = (1 - \lambda)\mu_X + \lambda X*$.....................(9)

where

$$\lambda = \frac{Var(X)}{Var(X*)} = \frac{Var(X)}{Var(X) + Var(U)} \Rightarrow 0 < \lambda < 1$$
……………………………..(10)

With a validation subset we can estimate

$$\hat{\lambda} = \frac{Va\hat{r}(X)}{Va\hat{r}(X*)}$$
…………………………………………………………… (11)

and

$$\hat{E}\left[X \mid X^*\right] = \left(1-\lambda\right)\hat{\bar{X}}^* + \hat{\lambda}X^* \quad \text{……………………………………………………. (12)}$$

$$E\left[Y \mid X\right] = \Pr\left(Y \mid X\right)E\left[Y \mid Y, X\right] \quad \text{…………………………………………….. (13)}$$

The first part Pr(Y|X) denotes the probability that a subject has mismeasued exposure given a set of variables X. The first part of the model is a weighted regression model.

The second part E[Y|Y,X] denotes the expected corrected mismeasured exposure Y given that the subject has corrected mismeasured exposure Y and a set of variables X. The second part of the modified two-part model is Bayesian regression model that will fit the data (Kamun, S. J. (2024)).

VI Bayesian Linear Regression

Modified Bayesian linear regression uses a weighted sum of variables to characterize parameter mean, aid in out-of-sample forecasting, determine prior distribution, and identify posterior distribution for model parameters.

The posterior expression is given below:

Posterior = (Likelihood * Prior)/Normalization

The formula calculates model parameters' prior probability based on the data's probability and posterior distribution, unlike OLS. As data accumulates, parameter values converge to OLS values, increasing accuracy.

In a linear model, if 'y' represents the expected value, then

$y(w,x) = w_0+w_1x_1+...+w_px_p$

where, the vector "w" is made up of the elements w_0, w_1,...w_p. The weight value is expressed as 'x'.

$w=(w_1\ldots w_p)$

As a result, the output "y" is now considered to be the Gaussian distribution around Xw for Bayesian Regression to produce a completely probabilistic model, as demonstrated below:

$p(y|X, w. \alpha) = N(y|Xw, \alpha)$

Where the Gamma distribution prior hyper-parameter alpha is present. It is handled as a probability calculated from the data (Kamun, S. J. (2024)).

VII Results

Small Sample Size Comparison and Selection.

By assessing the R squared values, bias, BIC, AIC, and standard error of rising sample sizes, from eight to twenty, the study looked at and examined these data.

Comparison of Sample Sizes

LARGE VARIANCE, ϵ ~ Gamma(0, 30)

Finding the Sample Size

Table 1: Summary of the R^2, RMSE, MAE, BIC, AIC,bias and standard error for sample sizes n from 10 to 20, with small variance "S", and with large variance "L"

Sample Size, n		NRMSE.mean. accuracy	**RMSE**	**MAE**	Multiple R-squared	Adjusted R-squared	**Bias**	Standard Error	**AIC**	**BIC**
8	L	0.99999987	5.9984 e-06	5.1512 e-06	0.999999999784	0.999999999497	1.8020 e-10	7.8034 e-11	-157. 6814	-157. 2047
9	L	0.93884012	3.1746	2.5522	0.723421568571	0.446843137142	0.1823	0.1156	58. 3341	59. 5175

10	L	0.99999985	6.9437 e-06	4.7558 e-06	0.999999999 746	0.999999999543	6.8407 e-11	3.2391 e-10	-197. 1746	-195. 3591
11	L	0.99999982	8.2502 e-06	6.9894 e-06	0.999999999 498	0.999999999164	2.1122 e-10	2.6328 e-10	-214. 2993	-211. 9119
12	L	0.9999998	9.0562 e-06	7.2897 e-06	0.999999999 704	0.999999999535	9.5158 e-11	2.2281 e-10	-232. 6349	-229. 7255
13	L	0.99999981	8.7633 e-06	7.7258 e-06	0.999999999 741	0.999999999611	6.3632 e-11	1.8516 e-10	-253. 8758	-250. 4861
14	L	0.99999983	7.6792 e-06	6.2851 e-06	0.999999999 772	0.999999999671	4.3298 e-11	1.3756 e-10	-278. 0258	-274. 1914
15	L	0.99999986	6.6324 e-06	5.9738 e-06	0.999999999 812	0.999999999737	3.8551 e-11	9.4696 e-11	-303. 1380	-298. 8897
16	L	0.99999985	6.9006 e-06	5.3446 e-06	0.999999999 809	0.99999999974	1.6044 e-11	1.4209 e-10	-322. 8788	-318. 2433
17	L	0.99999984	7.4587 e-06	6.2520 e-06	0.999999999 823	0.999999999764	2.9954 e-11	8.3149 e-11	-341. 1645	-336. 1652
18	L	0.99999981	8.7138 e-06	7.3317 e-06	0.999999999 701	0.999999999609	3.8156 e-11	1.5226 e-10	-356. 3398	-350. 9976
19	L	0.99999983	7.6726 e-06	6.1990 e-06	0.999999999 777	0.999999999713	2.1286 e-11	1.3157 e-10	-381. 6387	-375. 9721
20	L	0.99999982	8.3388 e-06	6.7944 e-06	0.999999999 761	0.999999999697	3.4073 e-11	1.0644 e-10	-399. 0263	-393. 0519

The study employed a sample size of n = 15, for high variance, ϵ ∼ N(0, 30). By carefully examining only the R^2 value, its accompanying bias, BIC, AIC, and standard error, these sample sizes best meet our criteria. So, for this study, a small sample of size n = 15 was utilized for large variance ϵ ∼ N(0, 30). Based on their R squared values, related bias, BIC, AIC, and standard error, Table 1 summarizes the results of the various small sample sizes.

The study observed appropriate values of the coefficient of determination the sample of small size with large error of size n = 15. For this investigation, the equivalent samples of modest sizes were sample with high error for n = 15. Based on the study selection criteria employed in Table 1 , the performance of the various sample sizes reveals that samples of n = 15 perform better than for other sample sizes.

Correction methods for measurement error

The study has employed the following five methods to correct measurement error in data from small samples: the Modified Two-part Bayesian regression methodology for correcting measurement error, the Simex Homoscedastic Measurement Error Correction Method, the Simex Heteroscedastic Measurement Error Correction Method, and the Systematic Measurement Error Correction Method. The study of their coefficient of determination R^2, bias, standard error, BIC, AIC, mean, and corresponding standard deviation of the corrected response variable forms the basis of our criterion for selecting the method for correcting measurement error. Table 2 provides results of summaries for small samples with high variance. Summary results of approaches for measurement error correction for small samples with significant error are shown in Table 2.

Table 2: Large Variance ϵ~N(0, 30)

Approaches for correcting Measurement Error		NRMSE	RMSE	MAE	R^2	bias	std.error	BIC	AIC	Mean	SD
Class.Meas. Err	L	0.9999998	8.338761e-06	6.794446e-06	9.99999999761067e-01	3.01427771631779e-11	1.10554889e-10	-393.0519	-399.0263	45.67436	0.5534801

Sys. Meas. Err	L	0.9976043	0.1096118	0.09241334	9.603521824e-01	0.004359288194	0.01939089235	-13.70049	-19.67488	45.67436	0.5647883
Simex Homo	L	0.9999998	1.105476e-05	9.064694e-06	9.99999999997e-01	3.95017352162e-13	1.79294660751e-12	-381.774	-387.7484	45.55808	6.062973
Simex Hete	L	0.9999997	1.250929e-05	9.993506e-06	0.99999999999623	4.78284079009e-13	1.91224489705e-12	-376.8296	-382.804	45.51853	6.608786
BLR	L	0.9999999	2.457741e-06	2.06485e-06	9.9999999999981e-01	3.6526337510168e-14	7.4323683703482e-14	-441.9188	-447.8932	45.69751	5.816396

According to the results of the methods for correcting measurement error based on coefficients of determination, sample bias, standard error, BIC, and AIC, all of the approaches seem to work well, with the exception of the systematic method for correcting measurement error, which exhibits lower coefficients of determination compared to the other methods, where one is a perfect fit and with relatively higher values for bias and standard error, BIC, and AIC, refer to Table 2 (0.892900855306). The selection criteria suggest that Simplex Homoscedastic Error, Simplex Heteroscedastic Error, and Bayesian Linear Regression techniques perform better, but the results of BLR are superior.

VIII Conclusion

In this paper, we brought out the comparison of the Modified Two-part Bayesian Regression with other existing approaches for measurement error. The comparison was made by using coefficient of regression, NRMSE, RMSE, MAE, bias, standard error, BIC, AIC, mean and standard error. All the above indicators show clearly that B.L.R performs better than most of the existing approach of Measurement Error Correction.

References

[1] Brazzale, A. R. and Guolo, A. (2008). A simulation-based comparison of techniques to correct for measurement error in matched case-control studies. Stat.med, vol. 27, issue 19, pp. 3755-3775.

[2] Carroll and Wand, M. P. Semiparametric estimation in logistic measurement error models. J. R. Statist. Soc. B, 53, 6, PP. 573-585, 1991.

[3] Rosner, B., Willett, W. C. and Spiegelman, D., (1992) Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Random Within-Person Measurement Error American Journal of Epidemiology, 136, 11, pp. 1400-1413.

[4] Breidt, F. J. and Opsomer, J. D. (2000), Local polynomial regression estimators in survey sampling, Annals of Statistics, 28, 1026-1053.

[5] Buonaccorsi, J. P. (2010). Measurement Error: Models, Methods and Application. Chapman Hall/CRC. [6] Buzas, J. S., Stefanski, L. A. and Tosteson, D. (2014). Measurement Error. In: Ahrens, W., Pigeot, I (eds). Handbook of Epidemiology. Springer, New York, NY. https://doi.org/10.1007/978-0-387-09834-0_19.

[7] Carroll, R. J., Ruppert, D. and Stefanski, L. A. (2006). Measurement Error in Nonlinear Models. Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781420010138.

[8] Dorfman, A. H. (2008), The two sample problem, Proceedings of the Joint Statistical Meetings, Section of Survey Research Methods. Journal of the American Statistical Association, 87, 998-1004.

[9] Fraser, G. E. and Stram, D. O. (2001). Regression Calibration in studies with correlated variables measured with error. Americal Journal of Epidemiology, vol. 154, issue 9, pp. 836-844.

[10] Freedman, L. S., Midhune, D., Carroll, R. J. and Kipnis, V. (2008). A Comparison of regression Calibration, Moment Reconstruction and imputation for adjusting for covariate measurement error in regression. Stat. Med. 27 (25): 5195- 5216; doi: 10.1002/sim3361.

[11] Kamun, S. J. (2024). The Modified Two-Part Bayesian Linear Regression Model for Estimating Error in Small Samples. International Journal of Research Publication and Reviews, Vol 5, no 6, pp 4437-4441

[12] Keogh, R. H. and White, I. R. (2014). A toolkit for Measurement Error correction, with focus on nutritional epidemiology. Stat.Med. 33 (12): 2135-55.

[13] Masser, K. and Natarajan, L. (2008). Maximum Likelihood, Multiple imputation and regression calibration for measurement error adjustment. Stat.Med. vol. 27, issue 30, Annual Conference of the International Society for Clinical Biostatistics, pp 6332-6350.

[14] Merkouris, T. (2004), Combining independent regression estimators from multiple surveys, Journal of the American Statistical Association, 99, 1131-1139.

[14] Rothman, K. J., Greenland, S. and Lash, T. L. (2008). Modern Epidemiology. Wolters Kluwer|Lippincott Williams & Williams.

[15] Spiegelman, D. (2013). Regression Calibration in air pollution Epidemiology with exposure estimated by spatio-temporal modelling. Environmetrics, 24 (8), 521. https://doi.org/10..1002/env.2249.

[16] THOMAS, d., Stram, D. and Dwyer, J. (1993). Exposure Measurement Error: Influence on Exposure-Disease relationships and Methods of correction. Annu. Rev. Publ. Health. 14; 69-93.