International Journal of Research Publication and Reviews

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

The Modified Two-Part Bayesian Linear Regression Model for Estimating Error in Small Samples

Samuel Joel Kamun

Catholic University of Eastern Africa
samuelkamun@gmail.com

DOI: https://doi.org/10.55248/gengpi.5.0624.1555

ABSTRACT

Using error correction techniques and ordinary least squares, the study created a modified two-part Bayesian linear regression model for estimating errors in small samples for response-selective data.

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

I. Introduction

Statistical problems involve measurement error, where variables cannot be accurately observed, often due to instrument and sampling errors.

II. Review of Literature

Agogo et al. (2014) and Brazzale et al. (2008) found that dietary errors in epidemiological research reduce the link between food intake and illness prevalence, affecting regression models.

Breidt et al. (2000) and Buonaccorsi (2010) explored local polynomial regression calibration estimators and measurement error situations in complex models, highlighting weaker assumptions and creative solutions for parameter estimation.

Buzas et al. (2014) and Carroll et al. (2006) discussed measurement error issues, which arise from self-reported data, suspect-quality records, biological, sampling, and analytical variability and require statistical models to fit unobservable variables.

Regression calibration prevents statistical power loss through calibration sub-studies and rudimentary measuring techniques by correcting biases in regression results brought on by imprecise quantification of exposure variables.

Epidemiological studies often face exposure measurement errors, leading to biased estimates of exposure-disease connections. Masser et al. (2008) explore three estimation approaches: maximum likelihood, multiple imputation, and regression calibration. ML performs better with substantial errors or large sample sizes.

Odile Sauzet et al. (2019) suggested a useful alternative of a two-part regression consisting of a logistic regression and a linear regression conditional on not being fully satisfied.

Bounthavong (2023) proposes a two-part model considering the large point mass of subjects with zero costs and non-parametric cost distribution properties, adjusting the data accordingly.

III. Methods

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

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

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.

III. Selecting and Comparing Small Sample Sizes

The study analyzed small sample sizes from eight to twenty, comparing R squared values, bias, BIC, AIC, and standard error, determining the appropriate sample size for the study.

IV. 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

$E[Y \mid X*, Z] = EX|X*,Z[E(Y \mid X*, Z) \mid X] = EX|X*,Z[E(Y \mid Z, X)] = EX|X*,Z[\beta_0 + \beta_X X + \beta_Z Z] = \beta_0 + \beta_X E[X \mid X*, Z] + \beta_Z Z$ ………….. (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∗):

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

where

$$\lambda = \frac{Var\left(X\right)}{Var\left(X^{*}\right)} = \frac{Var\left(X\right)}{Var\left(X\right)+Var\left(U\right)} \Rightarrow 0<\lambda<1$$

……………………………..(10)

With a validation subset we can estimate

$$\hat{\lambda} = \frac{Va\hat{r}\left(X\right)}{Va\hat{r}\left(X^{*}\right)}$$

………………………………………………………… (11)

and

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

…………………………………………………. (12)

$$E\left[Y \mid X\right] = \Pr\left(Y \mid X\right)E\left[Y \mid Y, X\right]$$

…………………………………………….. (13)

$$E\left[Y \mid X\right] = \Pr\left(Y \mid X\right)E\left[Y \mid Y, X\right]$$

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.

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.

V. Approaches for correcting Measurement Error

The study compared Modified Bayesian Linear Regression and OLS approaches for correcting measurement errors in small sample data, evaluating factors like R2, bias, standard error, BIC, AIC, mean, and standard deviation.

V. Results

The topic for this study is the Modified Two-Part Bayesian Regression Model for Estimating Error in Small Samples for Response-Selective Observations Using Multiple Regressions.

Finding the Sample Size

SMALL VARIANCE, $\epsilon \sim N(0, 1)$

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”.

Sample Size, n		NRMSE.mean. accuracy	**RMSE**	**MAE**	Multiple R-squared	Adjusted R-squared	**Bias**	Standard Error	**AIC**	**BIC**
8	S	0.99999994	2.3744 e-06	1.9684 e-06	0.999999998343	0.999999996134	1.1267 e-09	1.2791e-09	-172.5089	-172.0322
9	S	0.99999995	2.0147 e-06	1.7267 e-06	0.999999999035	0.99999999807	6.6648e-10	3.6578e-10	-198.5297	-197.3463
10	S	0.99999995	1.9168 e-06	1.4996 e-06	0.999999998892	0.999999998005	6.6809e-10	4.6161e-10	-222.9187	-221.1031
11	S	0.99999997	1.1849 e-06	1.0182 e-06	0.99999999966	0.999999999433	1.7289e-10	1.4412e-10	-256.9920	-254.6046
12	S	0.99999995	1.9306 e-06	1.4556 e-06	: 0.999999998293	0.999999997317	6.9446e-10	9.8263e-10	-269.7303	-266.8209
13	S	0.99999995	2.1353 e-06	1.701 e-06	0.999999998790	0.999999998185	-5.4341e-11	1.8460e-09	-290.5864	-287.1967
14	S	0.99999996	1.7601 e-06	1.2503 e-06	0.999999999299	0.999999998988	1.6024 e-10	3.9702e-10	-319.2729	-315.4386
15	S	0.99999993	2.7612 e-06	2.1200 e-06	0.999999997755	0.999999996857	3.3390 e-10	1.3912e-09	-329.4270	-325.1787
16	S	0.99999994	2.4863 e-06	1.7781e-06	0.999999998258	0.999999997624	3.0989 e-10	1.0173e-09	-355.5443	-350.9087
17	S	0.99999995	2.1379 e-06	1.6114 e-06	0.999999998764	0.999999998352	1.6269 e-10	6.7437e-10	-383.6490	-378.6497
18	S	0.99999995	2.0199 e-06	1.6270 e-06	0.999999999059	0.99999999877	1.3971 e-10	4.1842e-10	-408.96733	-403.6251
19	S	0.99999994	2.3297 e-06	4.1000 e-08	0.999999998702	0.999999998331	1.7693 e-10	6.0410e-10	-426.9308	-421.2642
20	S	0.99999994	2.5609 e-06	1.9559 e-06	0.999999998349	0.999999997908	1.6592 e-10	7.5697e-10	-446.2478	-440.2734

By using the metrics specified for this study, the results from Table 1 appear to suggest that the sample size of n = 11 is one that meets the requirements best. Hence, the study has used a sample size of n = 11.

Modified Bayesian Linear Regression (MBLR) and Ordinary Least Squares Regression

Table 2: Small Variance $\epsilon \sim N(0, 1)$, n = 11.

Approaches for correcting Measurement Error		NRMSE.mean. accuracy	RMSE	MAE	Coefficient of Determination R^2	bias	std. error	BIC	AIC	Mean	Standard Deviation
O.L.S	S	0.007949603	5.22233e-06	0.006828894	0.999999999994310	2.74302802694137e-12	2.74302802694137e-12	-237.4362	-239.82	53.53349	1.136945

M.B.L.R	S	0.9999999	2.886837e-06	2.424957e-06	0.999999999994909	3.91797705390218e-13	1.06130392046424e-12	-435.482	-441.4564	41.80354	1.874929

Notes:

O.L.S = Ordinary Least Squares Regression,

M.B.L.R =Two-Part Modified Bayesian Regression.

Just as was the case in Table 1, the metric used in Table 2 appears to suggest that M.B.L.R performs better than the O.L.S

The study compares a modified two-part Bayesian regression model for measurement error correction methods with OLS, revealing that M.B.L.R. outperforms O.L.S. based on the study's matrices.

VI. Conclusion

The study aims to find the best way to account for measurement error in small samples using the modified Bayesian regression model and compare it to OLS, another popular method, by looking at things like confidence measures, covariate variables, and coefficients of determination.

In the study, two methods for adjusting measurement errors were examined: Ordinary Least Squares (OLS) and Modified Two-Part Bayesian Regression Models. The modified two-part Bayesian regression approach demonstrated superior results based on selection criteria such as coefficient of determination, bias, standard error, mean, and standard deviation.

References

[1] Agogo, G. O., van der Voet, H., van't Veer, P., Ferrari, P., et al. (2014). Use of Two-Part Regression Calibration Model to Correct for Measurement Error in Episodically Consumed Foods in a Single-Replicate Study Design: EPIC Case Study. PLoS ONE 9 (11): e113160. doi: 10.1371/journal.pone.0113160. [2] 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.

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

[4] Buonaccorsi, J. P. (2010). Measurement Error: Models, Methods and Application. Chapman Hall/CRC. [5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] Mark Bounthavong (2023). Cost as a dependent variable: Application of the two-part model. Rpubs by Rstudios.

[11] 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.

[12] Odile Sauzet, Oliver Razum, Teresia Widera, Patrick Brzoska, (2023). Two-Part Models and Quantile Regression for the Analysis of Survey Data With a Spike. The Example of Satisfaction With Health Care. Epidemiology, Public Health, 11 June 2019, Volume 7 – 2019, https://doi.org/10.3389/fpubh.2019.00146